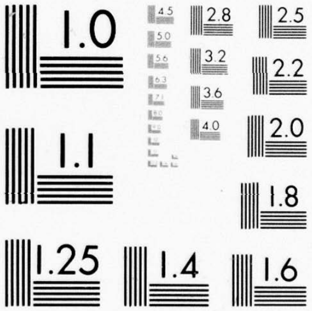
AD-A083 179 HASKINS LABS INC NEW HAVEN CONN F/G 5/10
SPEECH RESEARCH, A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
MAR 79 A S ABRAMSON, T BAER, C BEST, G CARDEN PHS-HD-01994
UNCLASSIFIED SR-57(1979) NL

1 OF 4
AD
A083179



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

*Status Report on*

# SPEECH RESEARCH

14 SR-57 (1979)

406643

9 Status Rep[redacted]t,[redacted] 1 Jan-31 Mar 79,

6 SPEECH RESEARCH,

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications,

12 296

11 1 January - 31 Mar[redacted]979

AD77663

10 Arthur S./Abramson Thomas/Baer
Catherine/Best Guy/Carden
Robert/Crowder

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

15 ✓ PHS-HD-~~0~~1994, PHS-HD-1-242~~0~~

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

406643 slt

## ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development
Grant HD-01994

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
Biomedical Research Support Grant RR-5596

National Science Foundation
Grant BNS76-82023

National Science Foundation
Grant MCS76-81034

National Institute of Neurological and Communicative Disorders and Stroke
Grant NS13870
Grant NS13617

National Institute of Arthritis, Metabolism, and Digestive Diseases
Grant AM25814

# HASKINS LABORATORIES

## Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper,* Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

### Investigators

Arthur S. Abramson*
Thomas Baer
Fredericka Bell-Berti+
Catherine Best+
Gloria J. Borden*
Guy Carden*
Robert Crowder*
William Ewan*
Carol A. Fowler*
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Alice Healy*
David Isenberg+
James J. Jenkins[3]
Leonard Katz[4]
Scott Kelso
Andrea G. Levitt*
Isabelle Y. Liberman*
Leigh Lisker*
Anders Löfqvist[2]
Virginia Mann+
Charles Marshall
Ignatius G. Mattingly*
Nancy McGarr*
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Donald P. Shankweiler*
Winifred Strange[3]
Michael Studdert-Kennedy*
Michael T. Turvey*
Robert Verbrugge*
Hirohide Yoshioka[1]

### Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Donald Hailey
Terry Halwes
Po-Chia Hsia*
Sabina D. Koroluk
Agnes M. McKeon
Nancy R. O'Brien
William P. Scully
Richard S. Sharkany
Leonard Szubowicz
Edward R. Wiley
David Zeichner

### Students*

David Dechovitz
Laurel Dent
Laurie Feldman
Hollis Fitch
Carole E. Gelfer
David Goodman
Janette Henderson
Charles Hoequist
Kenneth Holt
Robert Katz
Morey J. Kitzman
Peter Kugler
Roland Mandler
Leonard Mark
Suzi Pollack
Patti Jo Price
Brad Rakerd
Abigail Reilly
Arnold Shapiro
Louis G. Tassinary
Janet Titchener
Emily Tobey-Cullen
Betty Tuller
N. S. Viswanath
Douglas Whalen

*Part-time
[1] Visiting from University of Tokyo, Japan
[2] Visiting from Lund University, Sweden
[3] Visiting from University of Minnesota
[4] On sabbatical leave at the University of Sussex, U.K.
+NIH Research Fellow

# CONTENTS

## I. Manuscripts and Extended Reports

# I. MANUSCRIPTS AND EXTENDED REPORTS

# AN ARTICULATORY SYNTHESIZER FOR PERCEPTUAL RESEARCH*

Philip Rubin, Thomas Baer, and Paul Mermelstein+

Abstract. A software articulatory synthesizer, based upon a model developed by Mermelstein (1973), has been implemented at Haskins Laboratories. The synthesizer is being used as a tool for studying the linguistically and perceptually significant aspects of articulatory events. A prominent feature of this system is that it easily permits modification of a limited set of key parameters that control the positions of the major articulators: the lips, jaw, tongue body, tongue tip, velum and hyoid bone. Time-varying control over vocal-tract shape and nasal coupling is possible by a straightforward procedure that is similar to key-frame animation: critical vocal-tract configurations are specified along with excitation and timing information. Articulation then proceeds on a directed path between these key frames within the time-script specified by the user. Such a procedure permits the required degree of control over articulator positions and movements. The organization of this system and its present and future applications are discussed.

## INTRODUCTION

This paper provides a brief description of a software articulatory synthesizer implemented at Haskins Laboratories that is being used for a variety of experiments designed to explore the relationships between perception and production. A related paper, by Abramson, Nye, Henderson and Marshall (1979), describes in greater detail some of these experiments, focusing on the relationship between velar control and the distinction between oral stop consonants and their nasal counterparts. The intent of the present paper is to provide an overview of the actual design and operation of the synthesizer, with specific regard to its use as a tool for the perceptual evaluation of articulatory gestures.

The articulatory synthesizer embodies several sub-models. At its heart are simple models of six key articulators. The positions of these articula-

*Portions of this paper were presented at the 95th Meeting of the Acoustical Society of America, 16-19, May 1978, Providence, R. I.

+Bell-Northern Research and INRS-Telecommunications, University of Quebec, Montreal, Canada.

Acknowledgment: The research reported here was supported under NSF Grant BNS-76-82023 and BRSG Grant RR-05596 to Haskins Laboratories. We thank Steven B. Davis, Charles W. Marshall and Leonard Szubowicz for their significant contributions to the development of the articulatory synthesis program.

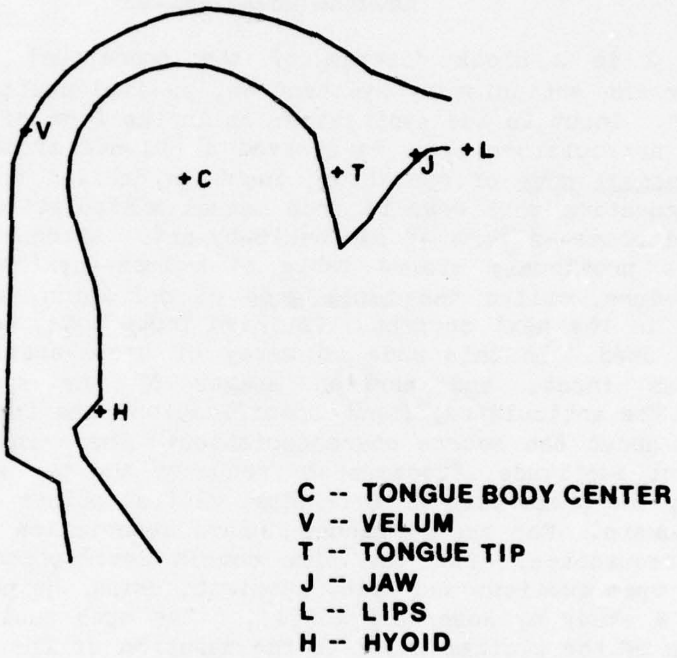
tors determine the outline of the vocal tract in the midsagittal plane. From this outline the width function and, subsequently, the area function of the vocal tract are determined. Source information is specified at the acoustic, rather than at the articulatory level, and is independent of the articulatory model. Speech output during each frame is obtained after calculating, for a particular vocal-tract shape, the acoustic transfer function for both the glottal and fricative sources. For voiced sounds, the transfer function accounts for both the oral and nasal branches of the vocal tract. Continuous speech is obtained by a technique similar to key-frame animation (see below).

Although the synthesizer is capable of producing short segments of quite natural speech with a parsimonious input specification, its primary application is not based on these characteristics. The most important aspect of the synthesizer's design is that the articulatory model, though simple, captures the essential ingredients of real articulation. Thus, synthetic speech may be produced in which articulator positions or relative timing of articulatory gestures are precisely and systematically controlled, and the resulting acoustic output may be subjected to both analytical and perceptual analyses. Real talkers cannot, in general, produce utterances with systematic variations of an isolated articulatory variable. Further, for at least some articulatory variables--for example, velar elevation and the corresponding degree of velar port opening--simple variations in the articulatory parameter produce complex acoustic effects. Thus, the synthesizer can be used to perform investigations impossible with real talkers, or investigations that are difficult, at best, using acoustic synthesis techniques.

## THE ARTICULATORY MODEL

The model that we are using was originally developed by Mermelstein (1973) and is designed to permit simple control over a selected set of key articulatory parameters. The particular set of parameters employed provides for an adequate description of the vocal-tract shape--while also incorporating both individual control over articulators and physiologically constrained interaction between articulators.

Figure 1 shows a midsagittal section of the vocal tract, with the six key articulators labeled: tongue body, velum, tongue tip, jaw, lips and hyoid bone position. (Hyoid bone position controls larynx height and pharynx width.) These articulators can be grouped into two major categories: primary articulators, whose movement is independent of other articulators (the jaw, velum and hyoid bone); and secondary articulators, whose positions are functions of the positions of other articulators (the tongue body, tongue tip and lips). The articulators of this second group all move relative to the jaw. In addition, the tongue tip moves relative to the tongue body. In this manner, individual gestures can be separated into components arising from the movement of several articulators. For example, the lip-opening gesture in the production of a /ba/ is a function of the movement of two articulators: the opening of the lips themselves, and the dropping of the jaw for the vowel articulation. Movements of the jaw and velum have one degree of freedom, all other articulators move with two degrees of freedom. Movement of the velum has two effects: It alters the shape of the oral branch of the vocal tract and, in addition, it modulates the size of the coupling port to the fixed nasal tract. This articulatory model is based, in large part, on knowledge of



KEY VOCAL TRACT PARAMETERS

Figure 1

the anatomy and physiology and systematic observations of X-ray data and accompanying acoustic recordings of natural utterances (Mermelstein, 1973).

Figure 1 also shows the graphical display provided by the synthesizer system that permits the user to simply modify the midsagittal vocal-tract shape. To accomplish this, the user selects one of the six key parameters, moves a set of cross-hair cursors to specify its new position, and the new vocal-tract outline is immediately calculated and displayed on the graphics terminal. Before discussing further aspects of user interaction with the synthesizer system, an overview of the program structure will be presented.

## PROGRAM ORGANIZATION

Figure 2 is a block diagram of the conceptual organization of the software for the articulatory synthesizer, as implemented on a DEC PDP-11/45 minicomputer. Input to the synthesizer is in the form of a list of positions of the key articulators that is arrived at in one of two ways. In what is called the *manual mode* of operation, input is derived from the static vocal-tract configuration that results from manual manipulation of the articulators as described above--a form of synthesis-by-art. Alternatively, input can be read from a previously stored table of values--synthesis-by-script. This second procedure, called the *table mode* of operation, will be discussed in more detail in the next section. (A third input mode, not shown in Figure 2, can also be used. In this mode an array of cross-sectional area values is specified as input, and earlier stages of the synthesis process are bypassed.) The articulatory input specifications are further supplemented by information about the source characteristics. These include, for the voiced source, input amplitude, fundamental frequency and two additional parameters that specify the properties of individual glottal pulses in either the time or frequency domain. For the frequency domain description these parameters are the pole frequencies. For the time domain description the two parameters specify the open quotient and speed quotient, using the pulse shape found most natural in a study by Rosenberg (1971). (The open quotient is the ratio of the duration of the glottal pulse to the duration of the whole glottal cycle. The speed quotient is the ratio of durations of the rising and falling phases of the pulse, and is thus a measure of skew.) In the case of fricative noise excitation, the amplitude and place of insertion of a pseudorandom noise source in the vocal tract are specified.

After the positions of the key articulators have been provided as input, either in manual or table mode, the program fleshes out this framework as a midsagittal section of the vocal tract, as was seen in Figure 1, and displays the shape, if desired. Cross-sectional areas are then calculated by superposing a grid structure on the vocal-tract outline and computing the intersecting points of the outline and the grid lines (Mermelstein, 1973). The center line of the vocal tract is determined as the line joining the mid-points of the grid segments subtended by the vocal-tract outline. This line length represents the length of the vocal tract modeled as a sequence of acoustic tubes of uniform cross-sectional areas. Vocal-tract widths are measured along lines perpendicular to the center line. Sagittal cross-sections are converted to cross-sectional areas with the aid of previously published data on the shape of the tract along its length. Different formulas are used for the pharyngeal region (Heinz & Stevens, 1964), oral region (Ladefoged, Anthony, & Riley,

STEPS IN ARTICULATORY SYNTHESIS

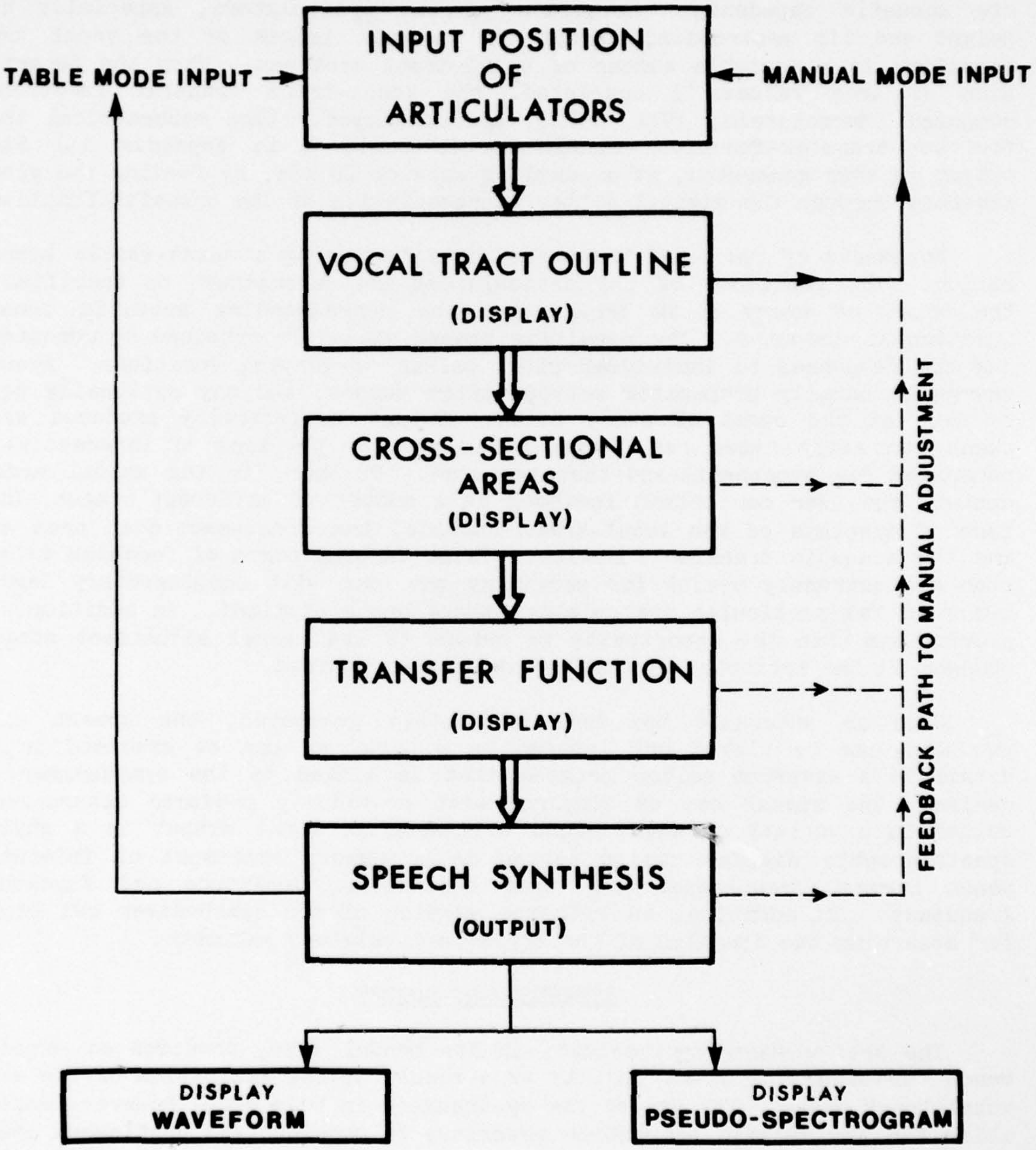


Figure 2

1971) and labial region (Mermelstein, Maeda, & Fujimura, 1971). The area function is then smoothed and approximated by a sequence of uniform tubes of fixed section length (.875 cm). To arrive at the cross-sectional area of the last section (nearest the lips), the tract is continued in a parabolic horn continuous in area with the computed cross-sectional area at the lips. This continuation allows truncation of the tract at a length value that is an integral multiple of the section length, and its termination by the appropriate acoustic impedance. Movements of the articulators, especially hyoid height and lip protrusion, affect the overall length of the vocal tract, resulting in a variable number of vocal-tract sections. When the determination of area values is completed, the vocal-tract transfer function is computed (Mermelstein, 1971, 1972), and displayed. (The mathematical theory for the transfer-function calculation is reviewed in Appendix A.) Speech output is then generated, at a sampling rate of 20 kHz, by feeding the glottal waveform through the digital filter representation of the transfer function.

Movements of the vocal tract are simulated using a quasi-static approximation. The positions of the articulators are determined, or specified, at the onset of every pitch period and the corresponding acoustic transfer function is computed. The resulting speech signal is obtained by concatenating the responses to individual pitch pulses of varying durations. Acoustic energy is usually propagated between pitch pulses, but may optionally be set to zero at the onset of every pulse. Output is generally produced within twenty to sixty times real time, which permits the kind of interactive use necessary for hypothesis-and-test research. Further, in the manual mode of control the user can obtain feedback at a number of different stages, in the form of displays of the vocal-tract outline, the cross-sectional area array and the acoustic transfer function. These varying forms of feedback information are extremely useful for providing the user with complementary descriptions of the particular articulatory shape being examined. In addition, they provide him with the opportunity to return to the manual adjustment stage if changes in the articulatory configuration are required.

Once an utterance has been completely generated, the speech signal produced can be played out, stored on a disk, or can be examined in more detail in a waveform editor program that is linked to the synthesizer. If desired, the signal can be compared with previously produced utterances or edited in a variety of ways. Also available as final output is a stylized spectrographic display, which serves as a summary statement of information about formant frequencies and their bandwidths, amplitude and fundamental frequency. In addition, an animated version of the synthesizer can be used for observing the dynamics of the entire articulatory sequence.

## SYNTHESIS-BY-SCRIPT

The articulatory synthesizer, in its manual mode, provides an excellent means for examining vowel quality as a result of the excitation of the static vocal-tract shape. The use of the synthesizer in this mode, however, does not allow for the dynamic simulation necessary to model actual continuous speech. Therefore, another procedure has been implemented to provide for time-varying control over the movements of the articulators. The approach used is similar to what is called key-frame animation: the framework for a desired dynamic articulation is represented by a series of configurations of the vocal tract.

The actual path of articulation is obtained by interpolating between these key frames. In essence, the user provides the synthesizer with a script, in the form of a table of values, for the complete articulation. Each line of the script consists of a "snapshot" of the vocal tract at some temporal point. The exact form of the articulation-over-time is then determined by linearly interpolating the articulatory parameters and computing the corresponding sequence of vocal-tract shapes.

An example of this procedure can be seen in Figure 3 for the synthesis of the utterance /da/. There are two "key" vocal-tract shapes specified as input. The first shape is an articulation appropriate for the onset of the production of the /da/, with the tongue tip occluding the vocal tract at the alveolar ridge. The vocal-tract shape appropriate for an /a/ is the second key configuration. The script, then, for this production begins with the first key shape specified at the onset of the utterance, at time 50 (ms), which permits a period of pre-release voicing. Release occurs at 50 ms, and all movement is completed by 120 ms, at which time the second key configuration is achieved. In this 70 ms period of rapid movement, a number of different intermediate shapes are calculated by linear interpolation, of which two (at 75 ms and 100 ms) are indicated in the figure. The production of the syllable continues for another 250 ms as specified by the time of the final /a/ configuration in the input. The additional specifications of this shape are necessary to indicate the changes in the excitation parameters, such as a fall-off in fundamental frequency towards the final third of the utterance (beginning at 240 ms), and a rapid decrease in the amplitude in the final 25 ms. The bottom half of Figure 3 shows the output generated from this articulation script in the form of a stylized spectrogram representing the first three formants, and the time-synchronized plots of fundamental frequency and overall output amplitude.

This straightforward procedure affords the user a flexible means of approximating productions observed in actual speech. Articulator movements are controlled by directing their path from shape to shape, with critical configurations serving as guides along the way. This allows for a simple specification of input information by the user. For example, the articulation for a /da/ can be represented, without refinements for naturalness, in the form of a script consisting merely of the two key vocal-tract shapes. Changes and comparisons between related articulations are easily accomplished. To produce a /na/ one can use the /da/ articulation previously described with the single modification of opening the velum to permit the required amount of nasalization. A series along the continuum from /da/ to /na/ can be created, then, by using ordered steps of velar opening, from a completely closed velum to one open to the degree desired for an acceptable /na/. Further, changes in timing relationships are also accomplished simply, by varying the time required to move between key configurations.

## CONCLUSION

The design and implementation of the articulatory synthesizer is intended, as previously noted, to provide researchers with a flexible interactive tool for examining relationships between speech perception and production. Input parameters to the synthesizer are the positions of a limited set of major articulators and excitation and timing information. An important aspect

# ARTICULATORY SYNTHESIS OF /da/

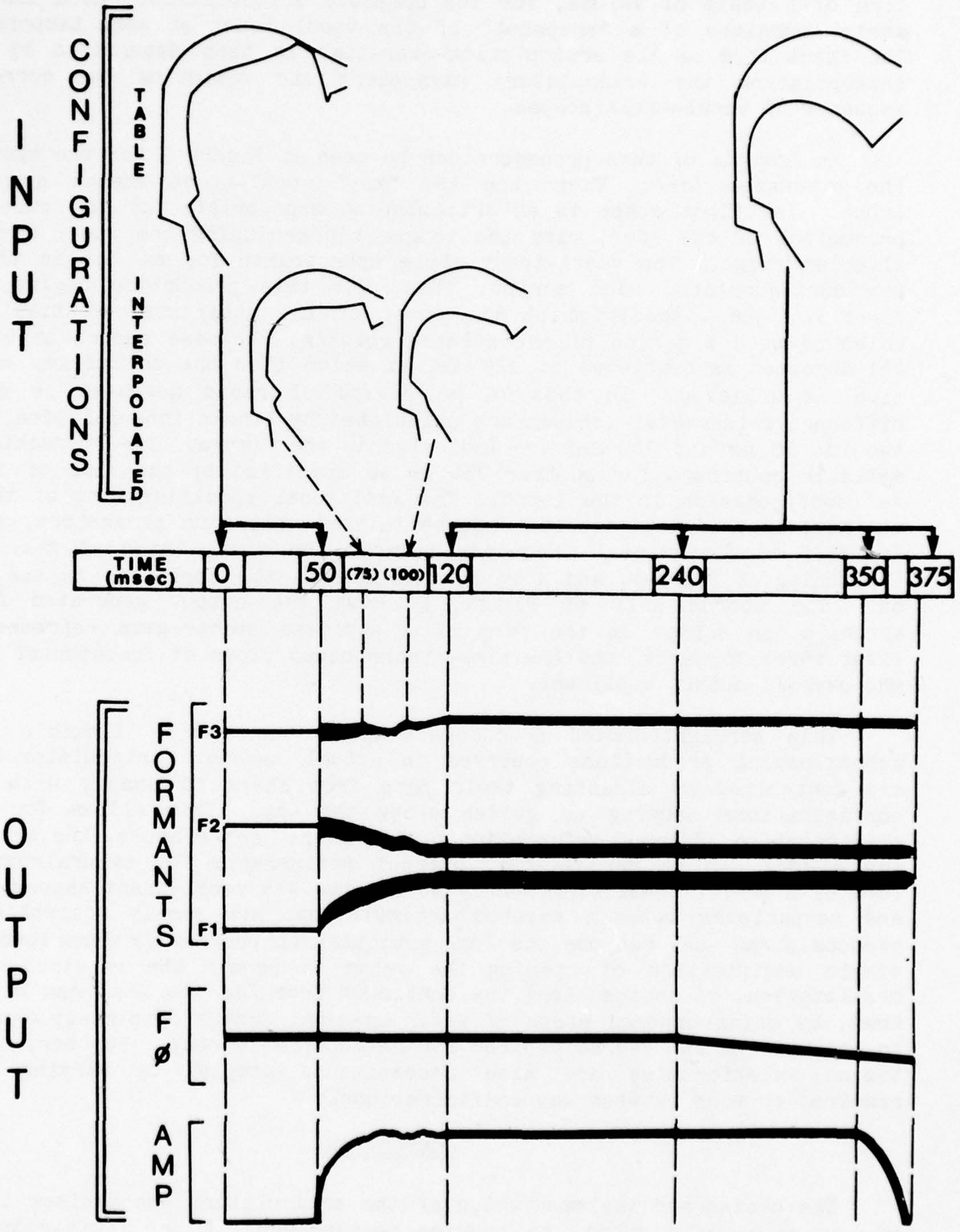


Figure 3

of the model's design is that speech sounds can be generated using controlled variations in timing and/or position parameters, and the output sounds used in formal perceptual tests. Another important aspect of the model is that the synthesis procedure is fast enough to make on-line interactive research practical.

One present application of the synthesizer is an investigation of detailed relationships between velar control and the perceptual oral-nasal distinction. Here, an important attribute of the synthesizer is its ability to produce complex variations in the acoustic output from a simple and natural variation of a single articulatory parameter--as contrasted with the more complicated procedures necessary to generate oral-nasal series by acoustic synthesis methods. In another application (Raphael, Bell-Berti, Collier, & Baer, in press), the articulatory synthesizer has been used to test hypotheses about articulation made on the basis of physiological (EMG) evidence on the one hand, and acoustic evidence on the other.

Additional future applications include a series of experiments intended to study the perceptual effects of variations in the relative timing of articulatory movements. Such investigations address the nature of the underlying organization of the speech act in terms of its dynamic "units." A planned technical improvement will be the development of a flexible display system that can function like a stop-frame projector. As we gain further insight into the anatomy and physiology of speech production, we would like to incorporate such additional knowledge into the model. The articulatory synthesis system, as described in this paper, already serves as a powerful research tool for examining perception - production relationships. We expect that the synthesizer's usefulness will grow as the system evolves and as we refine the issues to be investigated with its aid.

## APPENDIX A

### AREA-FUNCTION TO ACOUSTIC-TRANSFER-FUNCTION CALCULATION

The acoustic model of speech production, given the vocal tract area function, is indicated in Figure 4. The various branches of the vocal tract are treated as linear two-port networks. For voiced or aspirated sounds (Fig. 4a), where the velar port may be partially open, the glottal source $U_g$, feeds the left part of the pharyngeal branch. (For convenience, the glottal source impedance has been brought inside the box.) The right side of the pharyngeal branch is connected in parallel with the nasal and oral branches. On the right side of these two boxes appear the radiated nasal and oral pressure, each across an open circuit, since radiation characteristics have been brought inside the boxes. The output sound is the sum of the nasal and oral pressures. The junction point, at which the three subsystems are connected in parallel, corresponds anatomically to the top of the pharynx, at the level of the velopharyngeal port. However, if the nasal port is closed, the nasal branch drops out, and it no longer matters at what anatomical level the two remaining boxes split.

For fricative sounds (Fig. 4b), there is a noise source anterior to a constriction. The system splits into two parts: a front cavity, and a back cavity (which anatomically includes the constriction and also includes the

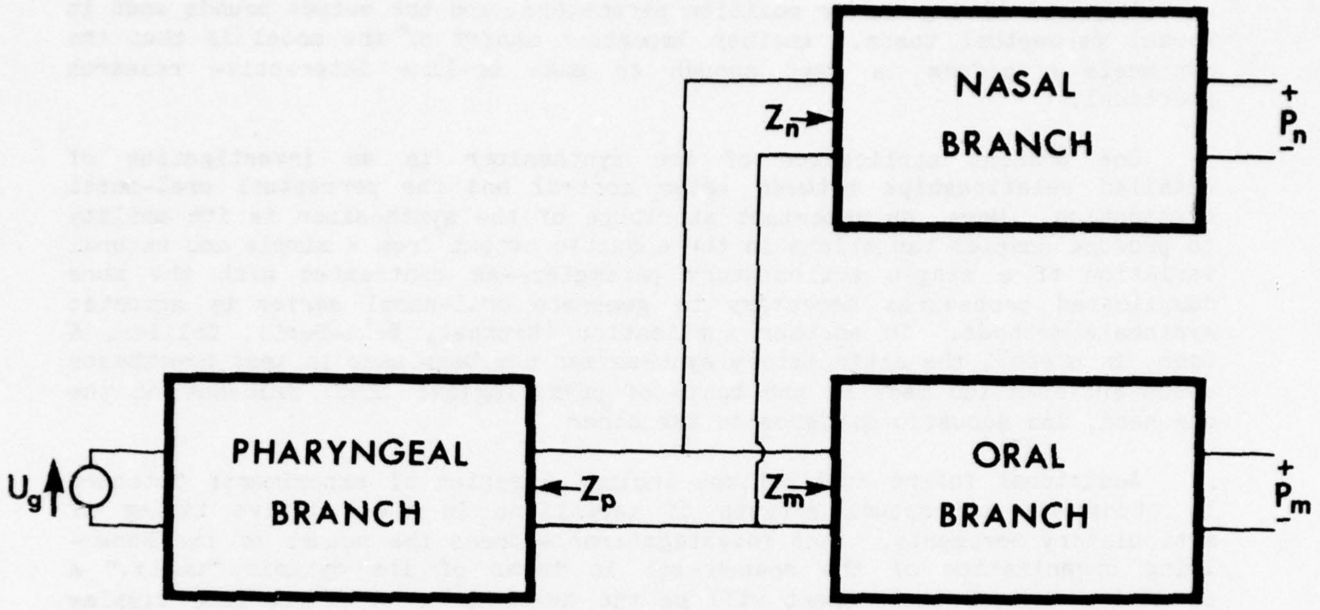


Figure 4a: Block diagram for voiced and aspirated sounds

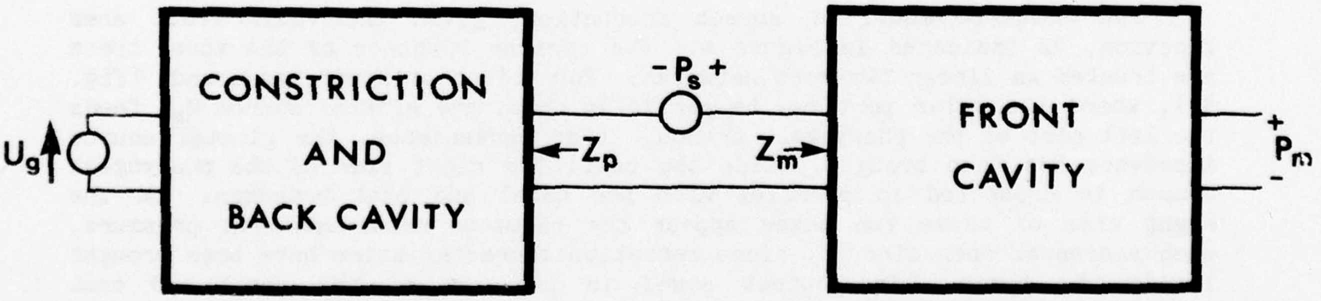


Figure 4b: Block diagram for fricatives

source resistance associated with the noise). Across the other side of the front cavity is the radiated pressure from the mouth, where, again, the radiation characterisitics have been brought inside the box. Across the other port of the back cavity appears the glottal source, if any. As before, glottal impedance has been brought inside the box.

In both parts of Figure 4, the glottal source and the leftmost box can be replaced by their Norton equivalent. If the two-port network obeys reciprocity, the Norton equivalent source, $U_{geff}$, is related to the actual glottal source, $U_g$, by the relation

$$U_{geff} = U_g G_p \ ,$$

where $G_p$ is the open-circuit pressure gain $p_g/p_p|_{Ug=0}$.

Using the Norton equivalent for the pharyngeal branch, as indicated in Fig. 5, it can be seen that the output for Fig. 4a is

$$p_m + p_n = \frac{U_{geff}}{1/Z_m + 1/Z_n + 1/Z_p} (G_m + G_n) \ ,$$

where $G_m$ and $G_n$ are the open-circuit gains $p_m/p_p|_{Ug=0}$ and $p_n/p_p|_{Un=0}$, respectively. Therefore the transfer function is

$$\frac{p_m + p_n}{U_g} = \frac{G_p \ (G_m + G_n)}{1/Z_m + 1/Z_n + 1/Z_p} \ . \qquad (1)$$

If the nasal tract is not present (that is, if the velopharyngeal port is closed), then $Z_n=\infty$ and $G_n=0$. Then, a corresponding equation accounts for the glottal component in Fig. 4b.

The transfer function for the fricative component in Fig. 4b is

$$\frac{p_m}{p_s} = \frac{Z_m}{Z_m + Z_p} G_m \ . \qquad (2)$$

Thus, all the relevant transfer functions can be calculated if the input impedances looking from the junction point and the open circuit pressure gain functions of all branches of the vocal tract are known. An iterative procedure for calculating these functions is described below.

For the purpose of calculations, the vocal tract is modeled as a series of uniform tubes with varying cross-dimension but uniform length of 0.875 cm. A plane wave entering one end of such a section reaches the other end with a half time-unit (0.025 msec.) delay and an attenuation $\alpha^{1/2}$, which depends on the cross sectional area. We will consider one such section with cross-sectional area A, looking into an acoustic impedance $Z_L$ from one end. When seen from inside the tube, this impedance produces a complex reflection coefficient

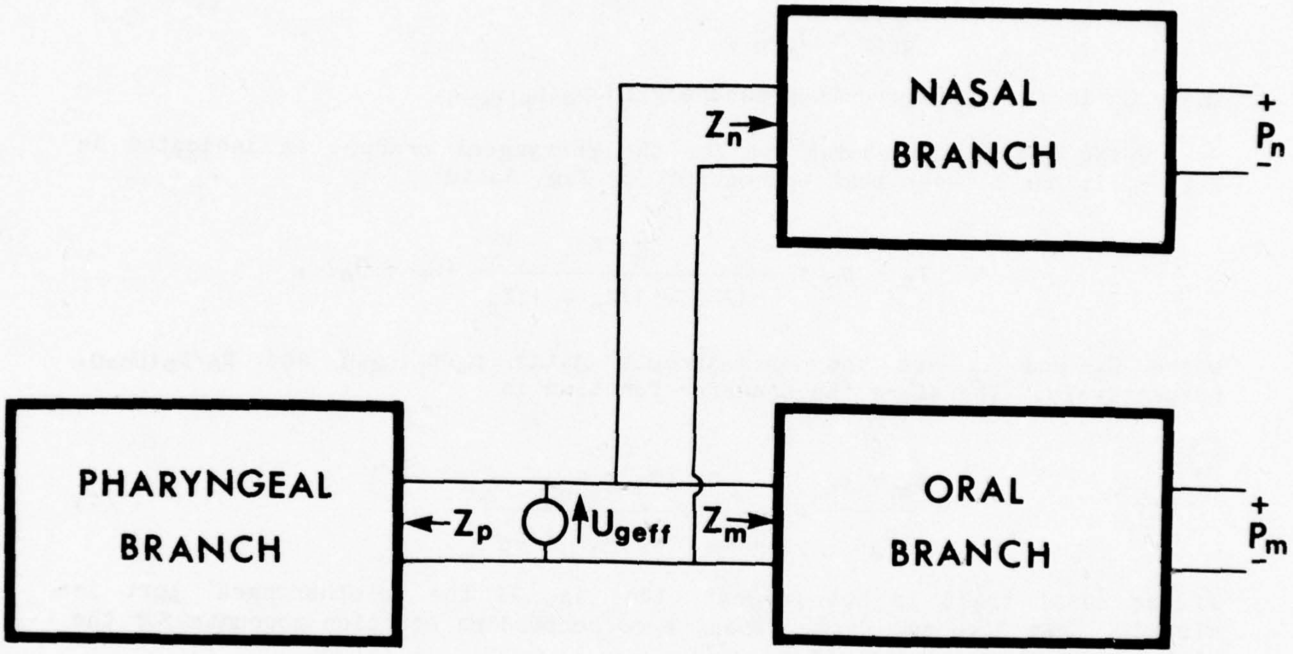


Figure 5: Equivalent block diagram for voiced and aspirated sounds

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} , \tag{3}$$

where $Z_0$=40/A is the characteristic impedance of the tube. (All physical quantities are expressed in cgs units.) The impedance, Z, looking into the other end of the tube is then

$$Z = Z_0 \frac{1 + z^{-1}\Gamma}{1 - z^{-1}\Gamma} ,$$

where z is the Z-transform variable and $\Gamma$ is expressed in Z-transform notation. The pressure gain across the tube is

$$G = (\alpha^{1/2} z^{-1/2}) \frac{1 + \Gamma}{1 + \alpha z^{-1}\Gamma} .$$

Consider now tube section n, of area $A_n$, looking into a load impedance $Z_{n-1}$, which produces the reflection coefficient $\Gamma_n$. The next section, which has area $A_{n+1}$, sees an acoustic impedance

$$Z_n = (40/A_n) \frac{1 + \alpha_n z^{-1}\Gamma_n}{1 - \alpha_n z^{-1}\Gamma_n} ,$$

which can be considered a reflection coefficient

$$\Gamma_{n+1} = \frac{r_n + \alpha_n z^{-1}\Gamma_n}{1 + r_n\alpha_n z^{-1}\Gamma_n} ,$$

where

$$r_n = \frac{A_{n+1} - A_n}{A_{n+1} + A_n} .$$

This can, in turn, be used to find the impedance or reflection coefficient on the other side of section n+1, and the gain across it.

We now express the reflection coefficients as ratios of polynomials, so that

$$\Gamma_n = P_n / Q_n ,$$

where P and Q are polynomials in z. Therefore,

$$P_{n+1} = r_n Q_n + \alpha_n z^{-1} P_n \tag{4}$$

and

$$Q_{n+1} = Q_n + r_n \alpha_n z^{-1} P_n . \tag{5}$$

The impedance into section n from the end of section n+1 is

$$Z_n = (40/A_{n+1}) \frac{Q_{n+1} + P_{n+1}}{Q_{n+1} - P_{n+1}} , \qquad (6)$$

and the pressure gain across section n is

$$G_n = (\alpha_n{}^{1/2} z^{-1/2}) \frac{Q_n + P_n}{Q_n + \alpha_n z^{-1} P_n} .$$

But

$$Q_n + \alpha_n z^{-1} P_n = \frac{Q_{n+1} + P_{n+1}}{1 + r_n} ,$$

so that

$$G_n = \alpha_n{}^{1/2} z^{-1/2} (1 + r_n) \frac{Q_n + P_n}{Q_{n+1} + P_{n+1}} ,$$

and the gain over sections 1 to N can be calculated:

$$G = z^{-N/2} \frac{Q_1 + P_1}{Q_{N+1} + P_{N+1}} \prod_{n=1}^{N} (\alpha_n{}^{1/2}(1 + r_n)) . \qquad (7)$$

To begin the transfer-function calculation, the source impedance or the radiation impedance (and gain) at the end of each branch of the vocal tract must be known and expressed as a ratio of polynomials in z. These are then used to determine the Q and P polynomials at the external end of the branch, using equation 3, and the iterative equations 4 and 5 are applied one section at a time until P and Q at the other (internal) end of the branch are determined. Lumped losses can also be introduced during these iterations. Both the impedance and gain can then be calculated, using equations 6 and 7, respectively. When this is done for all branches of the vocal tract, the glottal and fricative transfer functions can be calculated, using equations 1 and 2. Standard techniques are then used to implement the transfer functions as digital filters to perform the synthesis.

## REFERENCES

Abramson, A. S., Nye, P. W., Henderson, J., & Marshall, C. W. The perception of oral-nasal continua generated by articulatory synthesis. Haskins Laboratories Status Report on Speech Research, 1979, SR-57, this issue.

Heinz, J. M., & Stevens, K. N. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. Journal of the Acoustical Society of America, 1964, 36, 1037.

Ladefoged, P., Anthony, J., & Riley, D. Direct measurements of the vocal tract. *Journal of the Acoustical Society of America*, 1971, *49*, 104.

Mermelstein, P. Calculation of the vocal-tract transfer function for speech synthesis applications. In *Proceedings of the Seventh International Congress on Acoustics, Vol. 3*. Budapest: Akademiai Kiado, 173-176, 1971.

Mermelstein, P. Speech synthesis with the aid of a recursive filter approximating the transfer function of the nasalized vocal tract. In *Proceedings of the 1972 International Conference on Speech Communication and Processing*, Boston, Mass., 1972.

Mermelstein, P. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 1973, *53*, 1070-1082.

Mermelstein, P., Maeda, S., & Fujimura, O. Description of tongue and lip movement in a jaw-based coordinate system. *Journal of the Acoustical Society of America*, 1971, *49*, 104.

Raphael, L. J., Bell-Berti, F., Collier, R., & Baer, T. Tongue position in rounded and unrounded front vowel pairs. *Language and Speech*, in press.

Rosenberg, A. E. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 1971, *49*, 583-590.

# THE PERCEPTION OF AN ORAL-NASAL CONTINUUM GENERATED BY ARTICULATORY SYNTHESIS[*]

A. S. Abramson[+], P. W. Nye, J. Henderson[+] and C. W. Marshall

Abstract. Much evidence suggests that the perception of speech is intrinsically related to its production. By means of an articulatory synthesizer, the perception of the oral-nasal distinction in consonants was explored experimentally. This distinction was chosen because it is achieved by a very simple articulatory maneuver and because it is phonologically relevant in virtually every language. Lowering the velum in equal increments provided continua of CV syllables varying in size of velopharyngeal port which were divided perceptually into /d/ and /n/ categories by American English listeners. Discrimination experiments with a [da]-[na] series yielded the classical pattern of high discriminability at the category boundary, a result compatible with earlier observations of categorical perception. To test the hypothesis that lower (more open) vowels require a larger area of velopharyngeal coupling to give a nasal percept, three oral-nasal continua with the vowels /i/, /ʌ/ and /a/, respectively, were presented for identification. The results were in broad agreement with this hypothesis. The fact that the perceptual boundaries for the syllables employing /i/ and /ʌ/ were observed to be close to one another in velar port area, and significantly different from the larger velar port boundary associated with syllables incorporating /a/, also lent some support to the notion that identification boundaries are defined by a critical ratio of oral-to-nasal impedance.

## INTRODUCTION

With the aid of a computer model of the vocal tract designed by Mermelstein (1973) and improved by Rubin and Baer (1978), we have been reexamining certain speech phenomena that reveal evidence of the links between production and perception. In this paper we report on some of the perceptual effects arising from control of the velopharyngeal mechanism. The virtually universal phonological use of nasality to distinguish consonants, as well as its more limited use in vowels, makes it an important topic.

[*] This is a revised and expanded version of an oral paper presented at the 95th Meeting of the Acoustical Society of America, 16-19 May 1978, Providence, R.I.

[+] Also University of Connecticut, Storrs, Connecticut.

Acknowledgment: We wish to thank Thomas Baer for his thoughtful comments on this paper and Philip Rubin for his unstinting assistance in mastering the speech synthesizer. This research was supported under NSF Grant BNS-76-82023 and BRSG Grant RR-05596 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

We carried out two studies. The first explored some psychophysical aspects of discrimination behavior at the category boundary, and the second investigated a behavioral link between speech production and perception that determines the _position_ of the boundary between oral and nasal categories. In our exploration of this boundary positioning effect, we examined a theory that suggests that the boundary is determined by the relative impedances presented by the nasal and oral tracts.

The first of these studies was motivated by our curiosity concerning the underlying basis for arguments pertaining to categorical perception (Liberman, Harris, Hoffman & Griffith, 1957). These arguments rest upon the finding that listeners are largely unable to discriminate speech sounds that they would not normally label as different from one another. Thus, variants taken from a consonantal continuum can be sorted into phonological categories by listeners, but discrimination tests yield high levels of performance only in the region of the perceptual boundary on the stimulus dimension. The discrimination of variants within a single category is typically not much better than chance. We wondered whether the phenomenon of categorical speech perception might be a direct by-product of the fact that the intervals between successive synthesized samples from a continuum are customarily defined with reference to _spectral_ structure rather than to the structure of articulatory production. It could be argued that continua based on spectrally defined increments might give rise to the appearance of "nonlinear" perception (i.e., the perception would not change monotonically but would exhibit a discrimination peak at the category boundary) because a nonlinear relationship may exist between formant frequencies and displacements of the articulators. If the scale of the continuum were defined in articulatory terms and were thus based on a metric conceivably shared by the mechanisms of perception, "linear" discrimination perf rmance (i.e., discrimination functions without a peak) might be observed. We felt that our articulatory synthesizer now gave us the opportunity to test the plausibility of the linear versus nonlinear hypothesis, if we could select a case in which the specifications for articulatory production were relatively simple but the resulting output was spectrally quite complex and, therefore, difficult to synthesize accurately with any terminal analog synthesizer. The oral-nasal distinction met this criterion[1] since it is achieved in nature by merely lowering the velum and is reproduced with the articulatory synthesizer by manipulating the parameter that controls the area of the velopharyngeal port. Perhaps, if our hypothesis were correct, we would find that, with articulatorily defined utterances, our listeners would perceive linearly and fail to give us any discrimination peaks at the oral-nasal boundary.

Our second interest sprang out of the well-known correlation that exists between the size of the velopharyngeal port during nasal production and, in traditional terms, the height of the vowel: the higher the vowel, the higher the velum. Numerous observations made over more than forty years (Nusbaum, Foley & Wells, 1935; Harrington, 1944; and Bloomer, 1953; all cited in House & Stevens, 1956) have all attested to the vowel-height/velar-height relationship as a normal feature of speech production. In 1951, McDonald and Baker suggested that the correlation might be due to the speaker's efforts to maintain a "characteristic balance or ratio between oral and nasal resonance." This resonance ratio depends on the relative sizes of the velopharyngeal port and the posterior opening of the oral tract through the acoustic impedances that they present. Thus, when the speaker intends to produce no audible nasal output, a lower velum is tolerated for an open vowel than for a close vowel. Conversely, to achieve nasal excitation for a more open vowel, the speaker must lower the velum more than he would for a close vowel. Thus, the origin of this effect is usually presumed to be perceptual. However, it may also have consequences that span a broader context. Since coarticulation effects allied with perceptual phenomena operating both forward and backward in time are known to cause interactions between adjacent phones, it appeared likely that a similar interaction would be also observable in consonant-vowel syllables. For example, Bell-Berti, Baer and Niimi (1978) have shown that the effects of vowel height on velar height extend into adjacent consonants. Therefore, in the second part of this study, we set out to discover whether a correlation might also exist between the size of the velopharyngeal port required for _nasal consonant_ production and the height of the following vowel.

As far as we know, there is no anatomical or physiological reason to believe that velar activity is mechanically linked to motion of the tongue and jaw; hence, any variations in velar height cannot bear a _direct_ relationship to vocal-tract openness. We must suppose, therefore, that variations in velar height are the result of _voluntarily_ acquired habits and, hence, only _indirectly_ related to other articulatory events -- possibly through auditory feedback of the acoustic signal. Thus, if auditory monitoring is involved, we may hypothesize that the velopharyngeal port size corresponding to the boundary between an oral and a nasal _percept_, should be smaller for a close vowel than for an open vowel. A speech synthesizer based upon a model of the vocal tract offers the most efficient way of generating the utterances needed to test a hypothesis of this kind, since precise incremental control of the velum can be maintained throughout.

The first instrumental study of oral-nasal boundary phenomena was made by House and Stevens (1956) who, using an electrical analogue of the vocal tract, obtained perceptual data on velopharyngeal port size and vowel openness which were consistent with the observations of nasal articulation. In addition, their analysis tended to support the impedance-matching hypothesis of McDonald and Baker. However, House and Stevens' results depended on non-linguistic evaluations of _vowel_ nasality and involved unnaturally large amounts of oral-nasal coupling. The velopharyngeal port area required to achieve close to 100 percent nasal responses in House and Stevens' study was nearly 4 sq. cm. On the other hand, the observations of several investigators (Passavant, 1863; Björk, 1961; Nylén, 1961; Warren, 1967; Isshiki, Honjow & Morimoto, 1968) converge on the opinion that the active region of velopharyngeal control permits the aperture to range from zero to about 1 sq. cm. We felt that House and Stevens' results could be accepted with more confidence if new data supporting the hypothesis were obtained with utterances based on articulatory specifications that were physiologically more plausible and in a situation requiring linguistic judgments.

Using our articulatory synthesizer, we might have designed an experiment paralleling that of House and Stevens but have drawn our utterances from a language, such as Hindi or Portuguese, which has oral-nasal contrasts in high and low vowels. However, because we were interested in extending observations of the velar-height/vowel-height relationship and because we had begun work on speech discrimination, the first study mentioned in this paper, and had the utterances already available, we chose to work with consonants. Our approach involved examining an extension of McDonald and Baker's theory, namely, that speakers and listeners also adopt as their consonant category boundary a particular ratio of oral and nasal impedances. Thus, since the oral impedance is lower in open vowels, a nasal consonant coarticulated with such a vowel would require the speaker to adopt a lower position for his velum to ensure that the ratio of the impedances exceed some fixed oral/nasal category boundary value. Conversely, the theory suggests that the listener's decision as to whether a particular utterance is nasalized depends on whether he perceives that the impedance ratio within the speaker's vocal apparatus has matched the criterion value for a nasal utterance.

## METHODS

### Structure of Test Utterances

The utterances that we created to explore the two issues just raised were generated by a digital computer model of the vocal tract (Rubin & Baer, 1978; Rubin, Baer & Mermelstein, 1979). Control of the model was achieved by manipulating parameters that described the moment-by-moment positions in the mid-saggital plane of the tongue, lips, jaw, velum and hyoid bone. The model imposes a number of constraints upon the control parameters in order to make the configurations that it adopts conform, within certain limits, to normal vocal physiology. For example, two such basic features built into the model are: (1) the tongue is attached to the jaw so that jaw lowering also lowers the tongue, and (2) enlargement of the velar port narrows the dimensions of the oral tract in the velopharyngeal region. On the other hand, some constraints are the result of simplifications designed to speed up calculation of the speech output. For example, each vocal tract configuration is derived as a series of between 18 and 22 short cylinders of various cross-sectional areas connected end-to-end. To obtain the sound output, this compound tube is excited at its distal end by pulses representing glottal vibration and, when appropriate, by noise introduced at the "glottis" or any point of constriction along the tube. In addition, each vocal tract configuration is held constant during each "glottal" period.

The synthesizer specifications used to produce the oral-nasal series called for the velopharyngeal port to open (with glottal pulsing present and the oral tract occluded) for a period of 50 msec. prior to tongue tip release, and to maintain the nasal coupling throughout the vowel, giving a total duration for the utterance of 340 msec.[2] In the impedance experiment, three vowels were associated with the consonants /d/ and /n/. Irrespective of which vowel or velar port size specification was delivered to the synthesizer, the input excitation energy contour always remained the same. Thus, as a consequence of their higher radiation efficiency, the peak output energy delivered during the production of utterances incorporating open vowels was always higher than that observed in the utterances containing close vowels. Furthermore, the peak output level tended to be reduced by up to 3 dB as the nasal coupling increased.

Plots of the transfer function of each of the articulatory synthesizer's vocalizations were made: (1) midway through the 50 msec period prior to tongue tip release; (2) midway through the transition toward the steady state vowel and (3) at the point of peak vowel output. To illustrate the acoustic changes that occur, the transfer functions for the /i/ vowel environment are shown in Figure 1 at three points in time. In the upper half of Figure 1 are the spectra obtained with a closed velopharyngeal port; the lower half of Figure 1 contains the data for a (fully open) velar port of area 19.2 sq. mm. In each case, the transfer function of the vocal tract and the numerator of the transfer function are indicated by continuous lines and dotted lines respectively. The upper (non-nasalized) spectrum in Figure 1a indicates that the synthesizer output is effectively inaudible during the first 50 msec. Then the output energy rises rapidly (Figure 1b) during the transition period following tongue tip release, which is heard as the phoneme /d/, and eventually reaches the steady-state spectral configuration (Figure 1c) which is heard as /i/. At the other extreme, the lower (nasalized) spectrum in Figure 1d indicates that the pre-release output, or nasal murmur heard as the phoneme /n/, is considerably higher than its non-nasalized counterpart. The nasal murmur is followed, after tongue tip release, by a transition phase (Figure 1e) into the steady-state (nasalized) vowel shown in Figure 1f. Since the actual acoustic output spectrum is the product of the vocal tract transfer function and the spectrum of the "glottal" source, the spectrum of the source excitation has been included as Figure 2.

The synthetic utterances were delivered to panels of listeners over headphones (type TDH-39) in a quiet room at 80dB SPL measured at the peak of the vowel /a/. In some instances noted below, the levels of the close vowels were amplified to achieve a peak sound pressure level equal to that reached by the vowel /a/.

## Selection of an Incremental Scale

Before we were ready to address the questions we had set ourselves, we attempted to resolve a methodological issue concerning the choice of scale increments that were to separate our test utterances. The scale had to be a continuous function of velar port size and also had to yield data amenable to the analytic method that we planned to use to measure shifts at the category boundaries, namely, one based on the normal distribution.[3] We proposed to obtain identification data for different series of utterances based on different scales from which we could select the scale that yielded data falling closest to a normal sigmoid. However, in the final analysis, our tests for closeness of fit proved to be inconclusive and we were obliged to make our choice somewhat arbitrarily. Nevertheless, despite the unsuccessful outcome of the attempt, the choice of scale is of sufficient importance to deserve mention, since its significance may well extend beyond a methodological issue to the more fundamental question of whether it represents the basic dimension that is employed by both the speech production and perception mechanisms.

At the outset of our effort to find a suitable scale, three plausible possibilities presented themselves. Our scale could employ either equal increments of the synthesizer's velar port area parameter (i.e., a direct mapping, $\Delta$ A=constant, would achieve the desired sigmoid distribution at the boundary) or it could employ a transformation of the area parameter into equal increments of velar port radius[4] ($\Delta$ R=constant), or it could maintain a constant Weber fraction ($\Delta$ A/A=constant). The relationship between these three functions is shown in Figure 3.

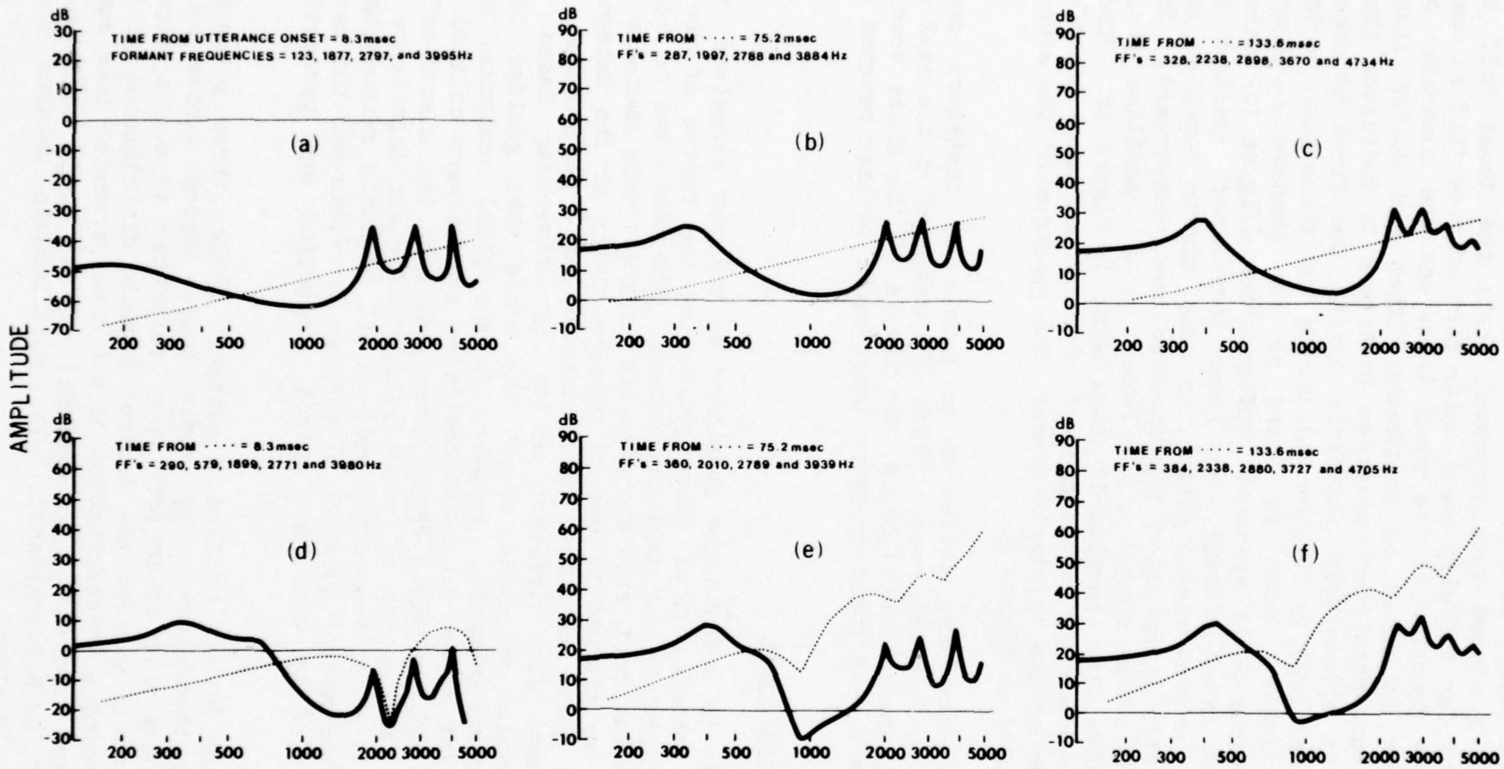


Figure 1: Vocal tract transfer functions at three points in [di] and [ni]. The three upper plots show transfer functions for a closed port, and the three lower plots, for an open port. Sections (a) and (d) occur 8.3 msec from the start of the 50-msec closure, sections (b) and (e) 75.2 msec after the release, and sections (c) and (f) in the steady-state vowel 133.6 msec from utterance onset. Dotted lines for the numerator of the transfer function show the spectral zeros. $F_0$ was 120 Hz.

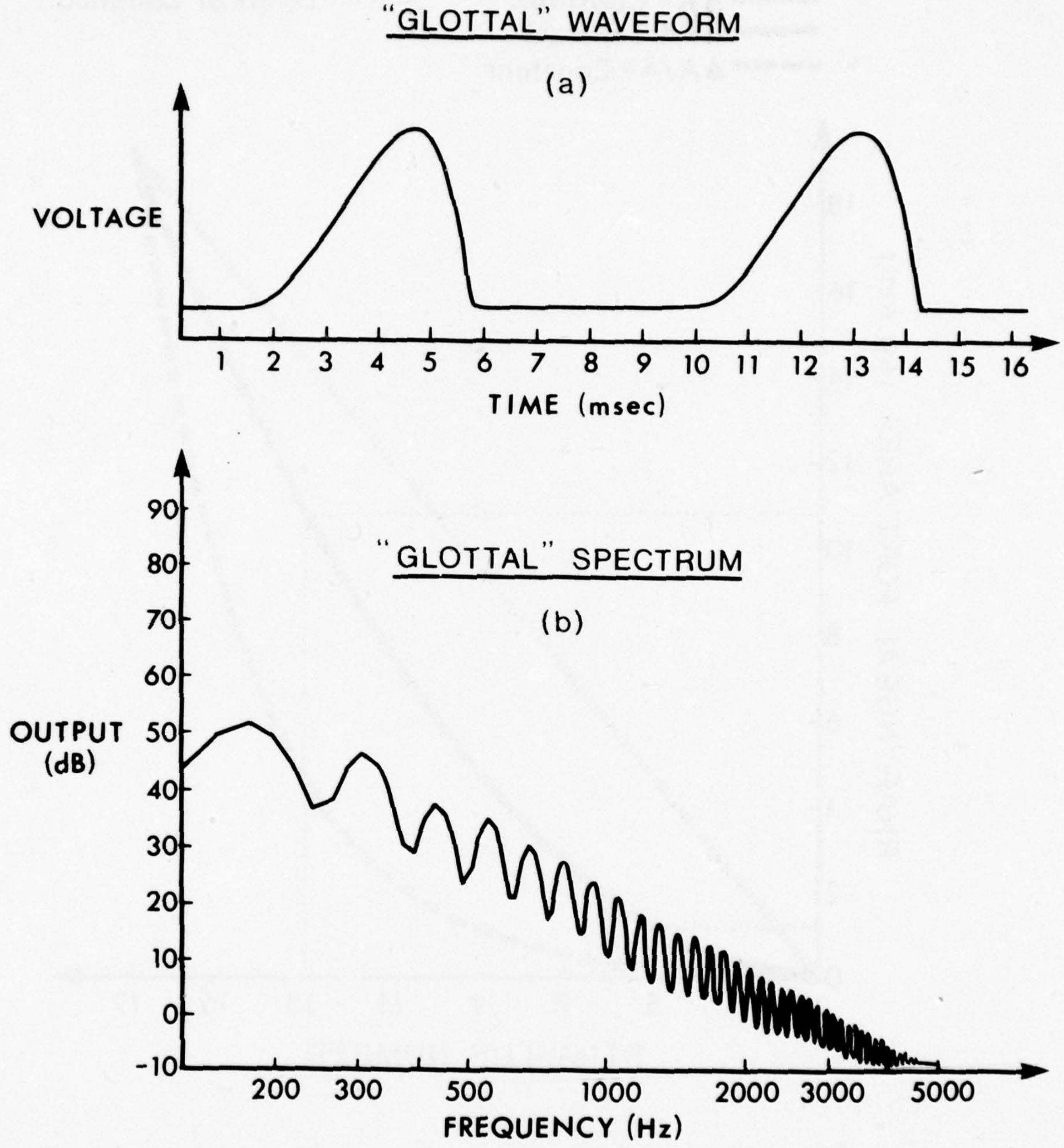


Figure 2: Section (a) shows the waveform of the "glottal" excitation used by the articulatory synthesis model and section (b) shows the spectrum of this waveform.

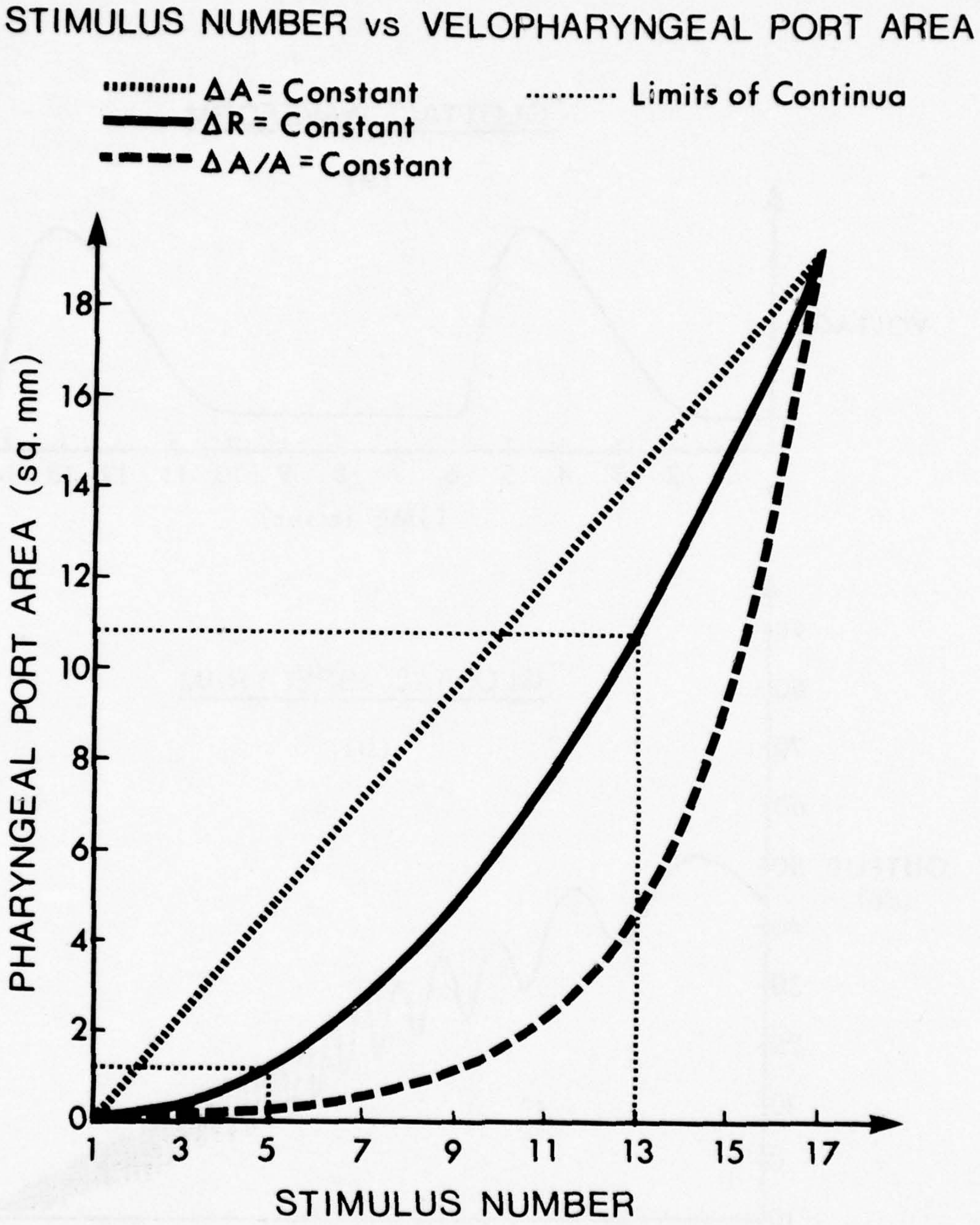


Figure 3: Relationship between three scale functions used to incrementally control velopharyngeal port size. Experiment 2 used 9 utterances which lay within the limits defined by the dotted line labeled "limits of continua."

The process of finding the stimulus scale began with informal listening tests of utterances lying in the interval between the English consonants /d/ and /n/ coarticulated with the vowel /a/. These tests were designed to find a nasal end-point with sufficient nasality to be accepted as [na] with the same confidence with which the extreme oral end with zero velar port area was accepted as [da]. The chosen velar opening for this nasal end-point was 19.2 sq. mm. For most of our listeners the response frequencies at the end-points reliably reached 100 percent and, with these end-points thus established, we began computing the first three series of seventeen velopharyngeal size specifications -- one advancing in equal increments of velar port area, another advancing by equal steps of velar port radius, and the third advancing in steps that maintained a constant Weber fraction. The resulting series of utterances for each type of increment was repeated 10 times, randomized and recorded on magnetic tape. The three tapes were then presented to 22 native speakers of English who heard them in one of three possible orders and identified the individual utterances as containing either the consonant /d/ or /n/.

To illustrate the method of analysis, Figure 4 shows the identification data obtained for a [da]-[na] series of utterances. A smooth curve was fitted to the percentage data using the Probit method of analysis (Finney, 1971) and the intercept of a 50 percent threshold line with the fitted curve was calculated on occasions when the category boundary was required. The intercept shown in Figure 4 is located at a velar port radius of 1.29 mm. (an area of 5.2 sq. mm.).

The outcome of fitting Probit curves to the three data sets from each individual listener and examining the Chi squared statistic for the goodness of fit was inconclusive. None of the three transformations yielded the smallest Chi squared for a significantly large number of listeners. The choice of the radius scale was ultimately made on the basis of the symmetrical position of its category boundary and the empirical evidence that the scale would respond no less accurately to the Probit method of calculating category boundaries than either of the other two scales.

## Experiment 1

Having determined that the radius scale could be employed without any relative loss of accuracy in the Probit analyses, we commenced our discrimination study and proceeded to create a nine-member continuum of "partially coarticulated" utterances from [da] to [na] that were equally spaced along the chosen continuum. The velopharyngeal port was completely closed at the oral end of the continuum and opened to 19.2 sq. mm. at the extreme nasal end. Discrimination tests were then prepared for one-, two- and three-step differences. The utterances were presented in AXB triads to groups of listeners such that either the first or the third member of the triad was physically identical with the second member; the listener's task was to say which utterance, the first or the third, was the same as the middle one. Each triad was repeated 10 times. For the one-step experiment, 25 listeners participated, while for the two- and three-step experiments, 18 listeners were employed of whom 12 were a subset of the original 25.

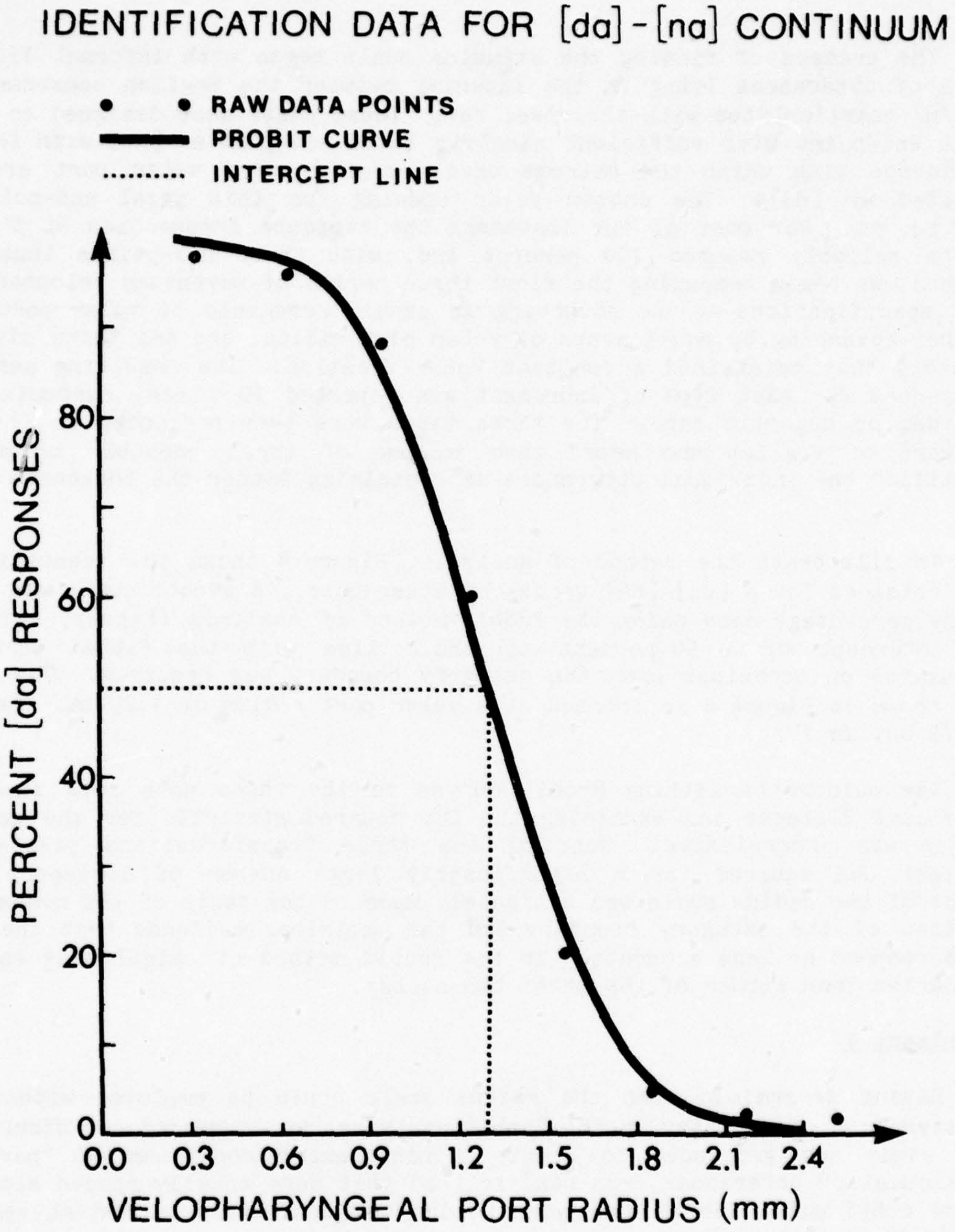


Figure 4: Identification data points, each representing the average score obtained from 25 listeners, are plotted for a continuum of utterances from [da] to [na] incremented along the radius scale. A fitted continuous curve has been computed by Probit analysis. A 50 percent threshold line intercepts the computed curve at a velar port radius of 1.29 mm.

Experiment 2

The study of impedance-matching effects employed the oral-nasal contrast in the English consonants /d/ and /n/ coarticulated with the vowels /i/, /ʌ/ and /a/. We chose articulatory configurations for the vowels /i/ and /a/ as end-points and then a configuration midway between these points for an acceptable /ʌ/. By means of the velar port size parameter, nine variants of velopharyngeal opening were computed along the radius scale. The velar port parameter at the oral end of the scale was set at the same value as the radius scale utterance No. 5 (velar port area of 1.2 sq. mm.) shown in Figure 3, and the extreme nasal end of the continuum was bounded by the radius scale utterance No. 13 (velar port area of 10.8 sq. mm.) from Figure 3. These two extrema plus the seven remaining intermediate values made up the nine velar port magnitudes specified for synthesizing the three continua based on the vowels /a/, /ʌ/ and /i/. The resulting utterances were recorded on magnetic tape under three different procedures and presented to native speakers of English for identification as the consonants /d/ or /n/.

The different recording procedures were employed in an effort to ensure that any observed boundary shifts were not a product of the method of stimulus presentation -- in particular that they were not a consequence of varying loudness levels and vowel-vowel interactions. Hence, in one procedure adopted for Experiment 2 (Condition 1), we employed tapes in which all three vowels were generated at their natural levels relative to the 80dB SPL standard level for the peak amplitude of /a/ and in which were mixed together three different randomizations of the utterances each containing 135 stimuli. In a second procedure (Condition 2), the randomizations of the three vowels were recorded on three separate tapes each of which contained 90 exemplars of a single vowel. On presentation of each tape, the output level of the recorder was adjusted to give an 80dB SPL peak vowel output. Finally, the third procedure (Condition 3) resembled Condition 1 with the exception that all the utterances containing the vowels /i/ and /ʌ/ were amplified prior to recording so that their peak vowel output levels were equated with the peak output of /a/. Hence, on replay, all the utterances contained vowels delivered at the same 80dB SPL peak level. It should be noted that under conditions in which the vowels were presented at their natural levels, the accompanying consonantal murmurs for any given velopharyngeal port size were reproduced at the same level for all three vowels. The procedure of equating peak vowel outputs disturbed this equality and resulted in progressively louder nasal murmurs for /a/, /ʌ/ and /i/ vowel contexts.

## RESULTS

### Linear vs. Nonlinear Discrimination

Evidence of categorical perceptual behavior is indicated when subjects find stimuli that fall within categories almost indistinguishable, while stimuli that straddle the category boundary can be discriminated with considerably greater frequency. The issue lay in whether the results of our identification and discrimination experiments performed with articulatorily synthesized oral and nasal consonants showed the familiar categorical behavior (i.e., "nonlinear" discrimination) or whether they exhibited a monotonically

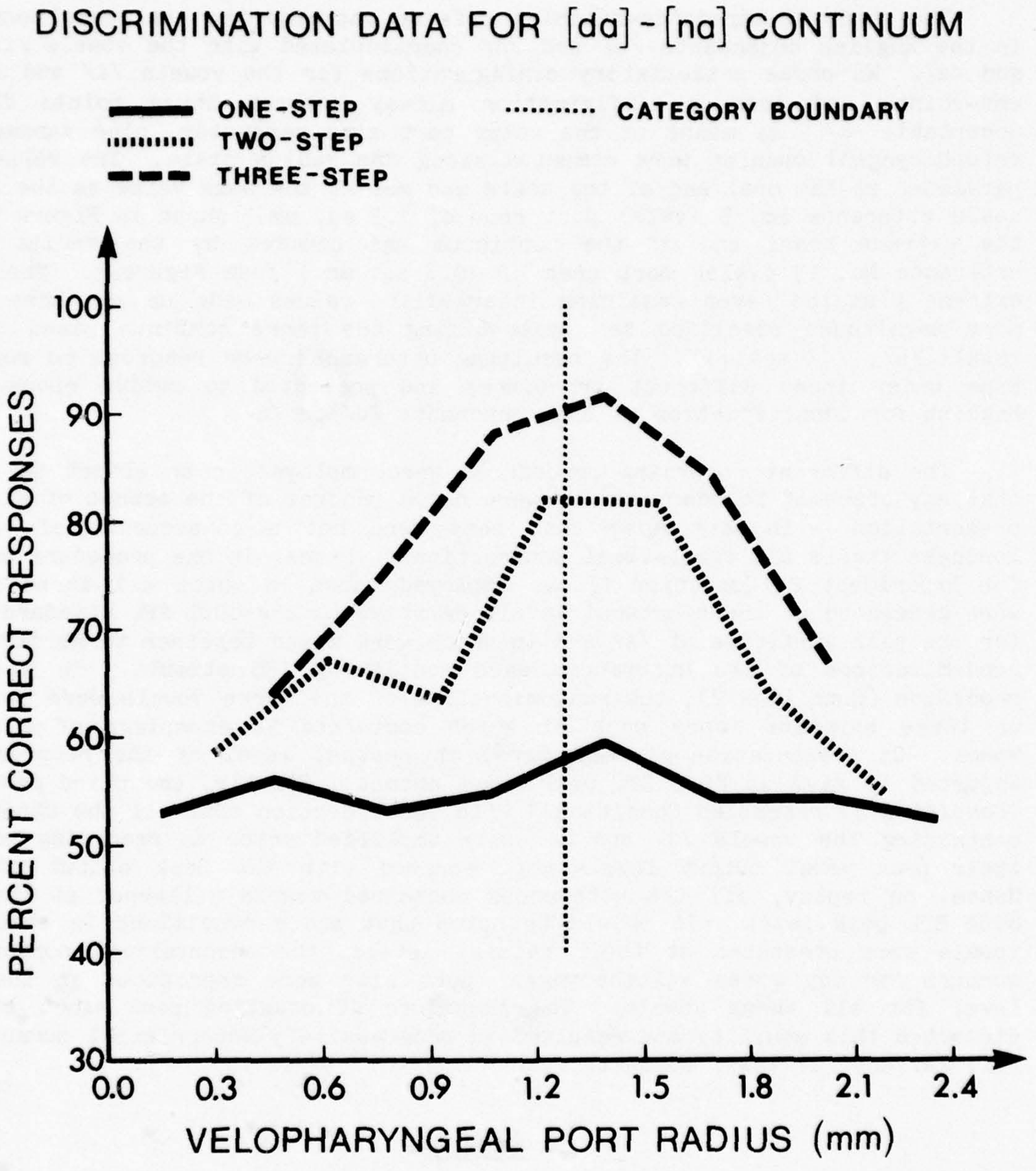


Figure 5: Discrimination data for a continuum of utterances extending from [da] to [na] incremented by units defined on the radius scale. One-step data are the averages for 25 listeners, while two- and three-step data are the averages for 18 listeners. The vertical line marks the oral-nasal category boundary obtained from an identification test (see Figure 4).

changing discrimination behavior across the category boundary (i.e., "linear" discrimination).

Identification data obtained with the [da]-[na] series of utterances composed for Experiment 1 are shown in Figure 4. The nine utterances are ordered along the abscissa from a closed velopharyngeal port on the left to the largest opening (a radius of 2.47 mm. or an area of 19.2 sq. mm.) on the right. The Probit curve is intercepted by the 50 percent threshold to give the category boundary at a velopharyngeal port radius of 1.29 mm. (an area of 5.2 sq. mm.). Figure 5 shows the companion discrimination data for the same series of utterances. The ordinate shows percentages of correct discrimination for the one-, two- and three-step differences as indicated by the coded lines. For the one-step difference, the listeners' performances hover around chance, but, for two-step and three-step differences, greater discrimination acuity emerges as a peak in the region of the category boundary. Thus, the discrimination functions of the listeners do not vary monotonically and the results do not appear to be in conflict with the classical observations. Moreover, the results are fully consistent with the nasal discrimination data obtained by other investigators who have used continua of spectrally specified increments in studies of oral-nasal perception (see Miller & Eimas, 1977).

## Relationship of Velar Height to Vowel Height

We now turn to the topic of the velar-height/vowel-height relationship and describe the results that we obtained from Experiment 2. We will begin with the Condition 1 procedure (mixed vowels at their natural output levels relative to /a/) in which a group of 14 listeners heard three randomizations of 135 utterances (9 values of port size x 3 vowels x 5 repetitions per randomization) and wrote down their consonant classifications. This group of listeners was formed from employees of the Laboratories who were not "naive" since all the listeners had gained some prior experience listening to synthetic speech, although very few had received much practice making repetitive quasi context-free linguistic judgments. The judgments provided by these listeners were tabulated as a function of velopharyngeal port size and vowel class, then expressed as percentages. Finally, Probit analysis was applied to find the intercepts of 50 percent threshold lines with the data.

In Figure 6a are plotted averages of the 14 intercepts (in units of velar port area; sq. mm.) for each of the three vowels in order of increasing openness. The graph indicates that the consonant category boundary does systematically move nasalward with the more open vowels and that the largest boundary shift occurs between the /i/ and /ʌ/ environments and the environment of the vowel /a/. This observation is confirmed by an analysis of variance showing that the probability that such a result could occur by chance is $p < 0.01$ $F(2,26)=5.62$, $MS_e=2.62$. Moreover, the Newman-Keuls test of significance applied to the differences between the boundary means shows the differences between the /i/ and /a/ and the /ʌ/ and /a/ environments to be significant at the $p < 0.05$ level. In other words, on the basis of these data, the case for a continuous movement of the category boundary with vowel openness was only moderately convincing. The procedure of Condition 2 was followed, therefore, to test the hypothesis further.

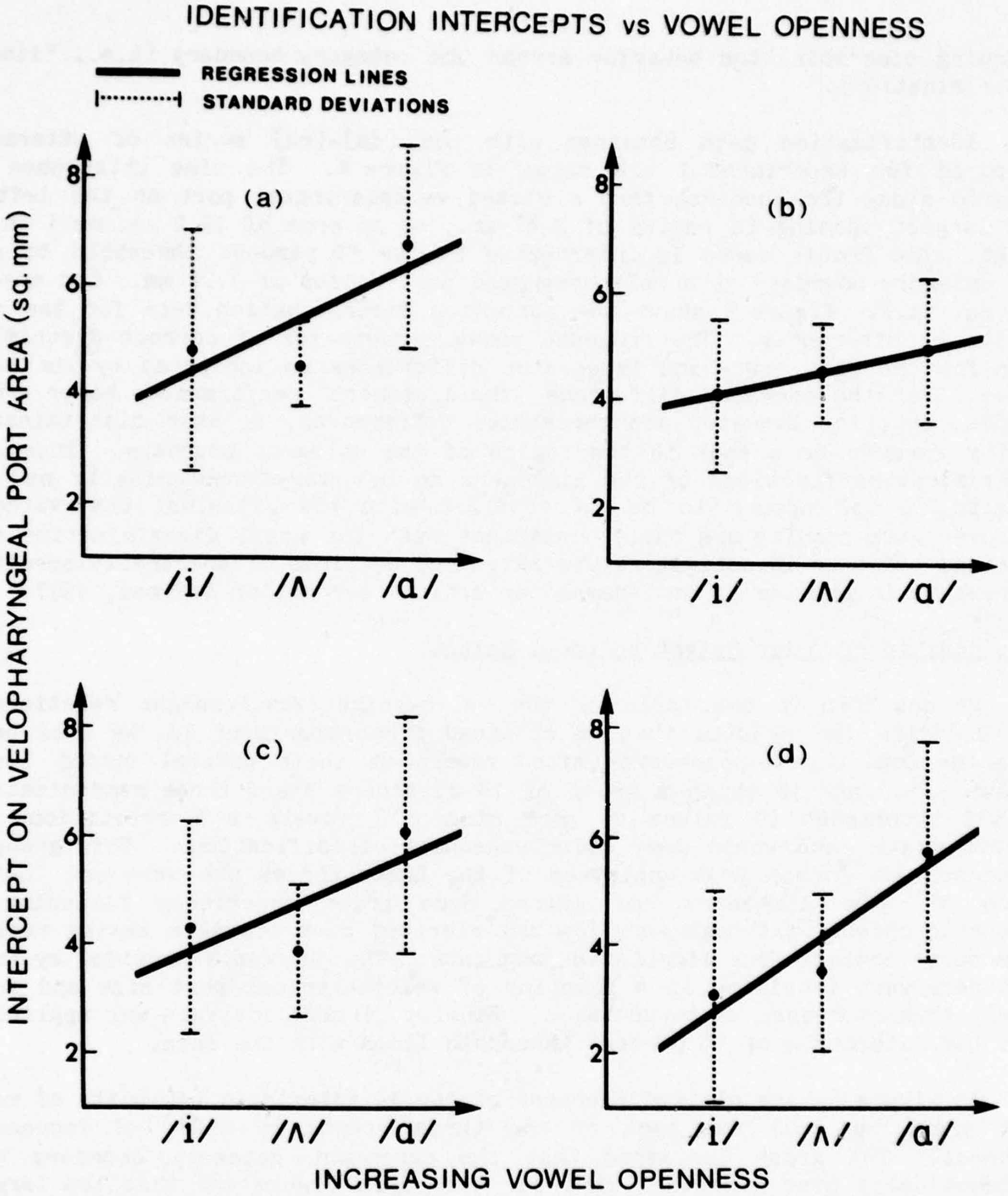


Figure 6: Results from four studies of the effects of vowel openness on the oral-nasal consonant boundary position. Average velopharyngeal port sizes at category boundaries are plotted against vowel openness with straight lines computed by linear regression. Broken vertical lines indicate standard deviations. Sections (a), (b), (c) and (d) were obtained under Conditions 1, 2, 3 and 3 respectively (see text).

In Condition 2, the recordings of the utterances were grouped under the three vowels and presented in three separate sessions. Each recording contained 10 repetitions at each of the 9 velar port sizes in randomized sequences. A panel of 24 listeners, drawn from the same source as before, was employed to identify the consonants and to write them down. The raw data obtained from these listeners, having first been subjected to Probit analysis as described earlier, yielded the graph shown in Figure 6b. Once again, a consonant boundary shift occurs in the direction indicating a greater tolerance of nasality in open vowels. While this boundary shift is somewhat smaller than the previous observation, the statistical evidence proves in fact to be stronger as a result of the lower mean-square error. In this case, the probability of the result's emerging by chance is $p < 0.01$ $F(2,48)=7.66$, $MS_e=0.519$ and the Newman-Keuls test shows the boundary difference between the /i/ and /a/ environment to be significant at the $p < 0.01$ level and the boundary difference between the /i/ and /ʌ/ environment to meet the $p < 0.05$ criterion. The /ʌ/ and /a/ boundary difference, on the other hand, does not meet a recognized criterion for significance.

The third variant of Experiment 2 (Condition 3, which employed mixed vowels adjusted to the same peak vowel output level) was performed on two groups of listeners; the first group of non-naive listeners consisted of 15 members of the panel employed in Conditions 1 and 2, while the second group consisted of experimentally naive students who carried out the listening task for pay. The results, shown in Figure 6c, repeat the earlier finding under Condition 1 in which there was a significant difference between the boundary locations in the context of /a/ and the contexts of /ʌ/ and /i/. The analysis of variance shows the boundary changes to be significant at the $p < 0.01$ level $F(2,28)=6.12$, $MS_e=2.27$ and the Newman-Keuls test reveals that the boundary differences between the /i/ and /a/ and the /ʌ/ and /a/ environments are significant at the $p < 0.05$ level. The data obtained under Condition 3 from the naive group of 17 listeners are plotted in Figure 6d. On this occasion, the difference between the category boundaries for /i/ and /ʌ/ environments is somewhat larger, while the category boundary for the /a/ environment is positioned at a velar aperture that comfortably exceeds that for both of the other vowels. The analysis of variance on these data yields $p < 0.01$ $F(2,32)=18.71$, $MS_e=1.84$.

Since the data obtained under all three conditions are broadly consistent, the accumulated evidence of these experiments is sufficient to justify the conclusion that listeners do tend to shift their consonant category boundary in such a direction as to require larger velar port apertures for nasal consonants when coarticulated with open vowels. Thus, they appear to compensate for the velar port size differences that are frequently observed in naturally produced vowels. It is also apparent that House and Stevens' boundary shift in the perception of vowel nasality, albeit more dramatic than ours based on consonant nasality, was not entirely an artifact of the unnaturally large amount of nasal coupling that they employed. However, on examining the average category boundary results contained in Table 1, there is one further question that emerges, namely, whether overall variations in vowel amplitude have any part in determining the position of the boundary.

TABLE I: Average Velar Port Area at Category Boundary (sq. mm.)

| | Vowel | Environments | | Overall |
|---|---|---|---|---|
| | /a/ | /^/ | /i/ | Difference |
| Condition 1: Mixed vowels, natural loudness. | 6.76 | 4.54 | 4.80 | 1.96 |
| Condition 2: Separate vowels, equal loudness. | 4.98 | 4.54 | 4.09 | 0.89 |
| Condition 3: Mixed vowels, equal loudness. | 6.00 | 3.88 | 4.29 | 1.77 |
| | 5.75 | 3.50 | 3.07 | 2.68 |

## Loudness Effects on Category Boundary

An examination of the first three rows of boundary values, which were all obtained from non-naive listeners, shows that Condition 1 gave rise to the largest boundary difference between /a/ and the other two vowels. Since this condition was the only one in which the overall vowel amplitude was allowed to vary, it was necessary to determine whether peak vowel amplitude could affect the position of the boundary. Therefore, we ran one additional experiment which, on this occasion, employed 17 naive listeners who heard the [di]-[ni] tape from Condition 2 played twice, once with the vowel peak output level set at 75 dB SPL and for a second time at a level of 90 dB SPL. Eight of the subjects heard the tape played with its sound pressure levels set in the opposing order.

The results showed a small but statistically insignificant shift in the position of the boundary. Thus, the magnitude of any boundary shift due to loudness level is clearly insufficient to account for the oral-nasal category boundary difference between open and close vowels. On the other hand, as the data in Table 1 show, procedural differences in presenting the utterances do affect the absolute boundary locations although they leave their relative relationship essentially unchanged.

## Impedance-Matching Hypothesis

At this juncture, there is the remaining question of whether McDonald and Baker were right in attributing the boundary shift effect to the perceiver's use of a balancing point between oral and nasal impedances as the position of his category boundary. In order to reexamine this idea, we computed impedance relationships for our vocal tract configurations to discover whether a constant impedance criterion could be found that would predict the relationship between the boundary locations we had obtained for our chosen vowels.

House and Stevens (1956, p.222) pointed out in their study of vowels (which touched upon the McDonald and Baker hypothesis) that the nose radiates much less sound energy than the mouth simply because it is smaller and more heavily damped. Hence, the principal effect of the nasal coupling is to modify the output of the oral tract. If, however, the impedance of the nasal tract is high in relation to that of the oral tract, the effect of nasal coupling on the output of the mouth is small. Conversely, a low nasal impedance results in a strong effect of coupling on the oral output. The concept of an impedance ratio is complicated, however, by the nature of the oral and nasal impedances. These fluctuate as a function of frequency and form a ratio that also varies with frequency as well as with the degree of nasal coupling. Thus, there is no single impedance ratio and the question arises as to which frequency, or band of frequencies, might be considered to be chiefly involved in the perception of the oral-nasal boundary and from which a "perceived impedance ratio" might be derived. House and Stevens tackled this problem by arguing that the changes in the vocal output are most likely to occur in those spectral regions where the difference between the oral and nasal impedances is greatest (i.e., where the ratio of oral impedance to nasal impedance is a maximum) and where output signal energy is usually highest (i.e., in the region of the first formant, $F_1$). This argument suggests that for a given amount of nasal coupling, the $F_1$ output of a close vowel should be modified more than the output of an open vowel like /a/. Data gathered by these authors on the relationship between velar port area and the relative amplitude of $F_1$ did indeed reveal larger perturbations for close vowels. Moreover, when these data were combined with perceptual data on the relationship between velar port area and vowel nasality for vowels of varying openness, House and Stevens were able to provide indirect[5] support for the impedance-matching hypothesis by showing that the relative amplitude of $F_1$ at the category boundary remains essentially constant as a function of vowel openness.

Based on a rationale similar to that proposed by House and Stevens, the "perceived impedance ratio" (a mean of the frequency-dependent ratio weighted by the $F_1$ amplitude within a range of 200Hz on either side of the formant peak) is plotted in Figure 7 as a function of velar port size for the three vowel environments used in the experiments. It is apparent that for any given impedance ratio criterion that intercepts all three vowel impedance functions within the utterance range (e.g., assume an arbitrarily chosen ratio of 0.2), the projection of those intercepts onto the abscissa of velar port size does not predict the observed oral-nasal category boundaries listed in Table 1; however, the impedance functions do exhibit one salient feature of the observations, namely, that the intercepts on the abscissa indicate the distance between the /i/ and /ʌ/ environments to be less than their separation from the /a/ vowel environment. In this respect, the results of the impedance analysis can be said to provide some (albeit qualitative) support for the impedance-matching hypothesis.[6]

## DISCUSSION

Our newly available techniques of articulatory synthesis of speech have been shown to be powerful enough to yield continua of variants which native speakers of English have no difficulty in dividing into categories of oral and nasal consonants. That is, incrementally increasing the size of the velopharyngeal port of our model brought about suitable acoustic coupling between the oral and nasal cavities which caused listeners to shift their labeling behavior from a well-established oral category through a brief zone of ambiguity to a well-established nasal category.

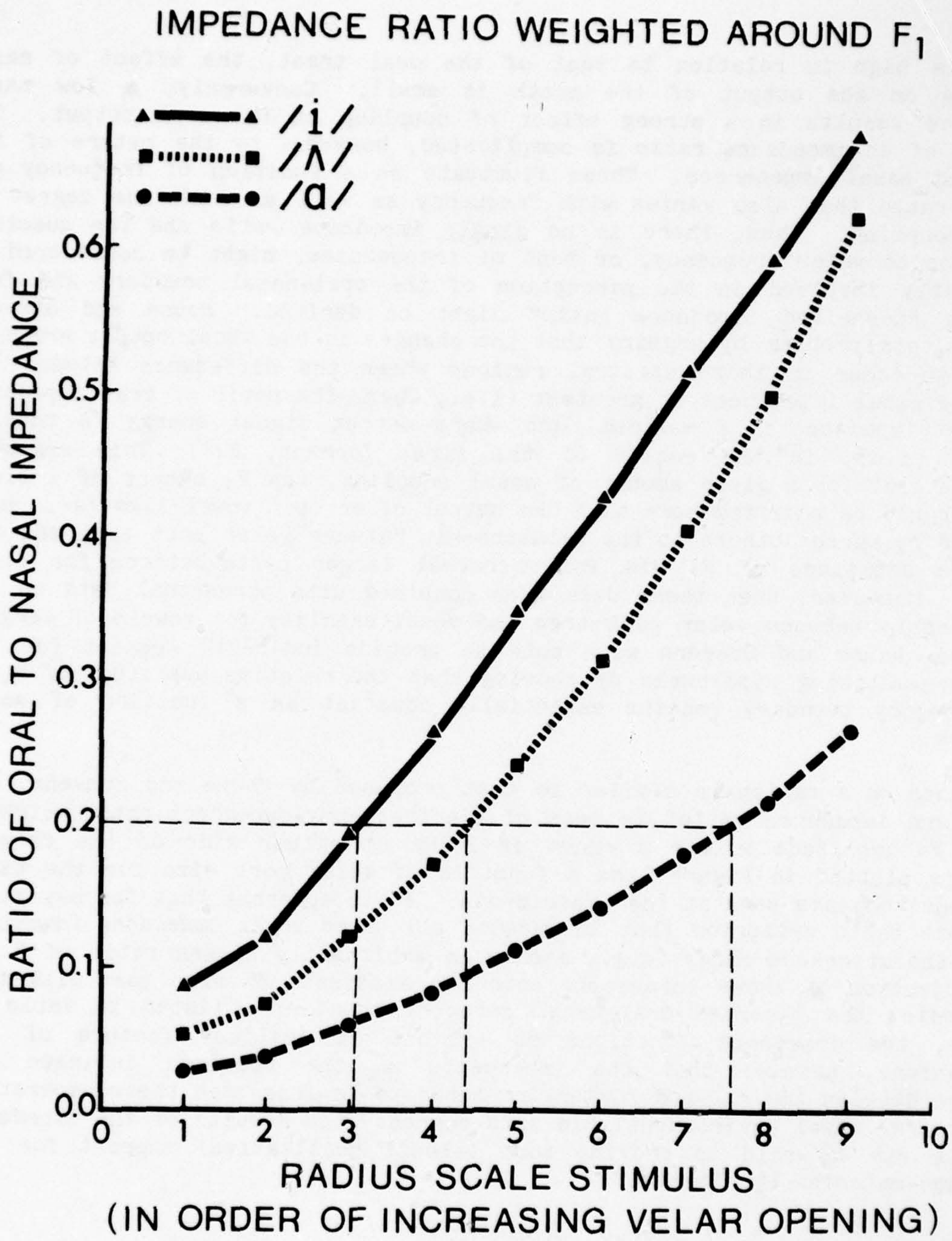


Figure 7: "Perceived impedance ratio " (see text for definition) is plotted as a function of velar port aperture for the three vowel environments /i/, /ʌ/ and /a/. An arbitrary threshold level intersecting the three functions leads to intercepts on the axis of velar port size.

In the course of our research, we were confronted with the problem of how best to set the intervals for our continuum of oral-nasal variants. What scale of increments, we asked, would be most compatible with the perceptual processing of variants along this phonetic dimension? In our articulatory synthesizer, we can directly effect changes in the area of the nasal coupling by varying the parameter controlling the velopharyngeal aperture. Therefore, we used a scale of equal increments in area and, in addition, we calculated scale values of velar port area that would provide us with equal increments in port radius and a constant Weber fraction. Although, as we have argued above, this methodological question seemed important enough to discuss here, our results were not conclusive, so we were obliged to make our choice of the radius scale on the basis of statistically inadequate data.[7]

Our foray into research on categorical perception was meant to examine the possibility that the classical findings of high acuity of discrimination in the region of the boundary between phonetic categories were artifactual because of nonlinearities between the acoustic continua used and articulation. Using our articulatory synthesizer, however, to make a continuum from [da] to [na] in English, we obtained the normal results in such discrimination tests. While not wishing to enter into any controversy concerning the reasons for categoricalness in speech perception and its possible links to articulatory control, we can at least conclude that letting the output spectrum of our synthetic syllables vary with increments of articulatory change, i.e., changes in the size of the velopharyngeal port, yields results that are in agreement with recent experiments using a terminal analogue synthesizer on the same phonetic distinction.

In connection with the dependence of velar port size at the oral-nasal category boundary on vowel openness, we have extended the oral-nasal boundary observations from work based exclusively on vowels (House & Stevens, 1956) to circumstances where nasal consonants are coarticulated with vowels of three different heights. Our results confirm that a significant shift in category boundary occurs. We have also examined the impedance-matching hypothesis of McDonald and Baker (1951) which proposes that the category boundary is established by the speaker-hearer as being some constant ratio of oral-to-nasal impedance. In this case our results have lent some support to the hypothesis since the identification data are in qualitative agreement with the perceived distances between the category boundaries as obtained by an impedance ratio analysis. From both of these studies, the results we have obtained have been somewhat less striking than those of House and Stevens, but we think that there are three reasons for this difference. First, they experimented with vowels, whereas we used consonants. Our test of the hypothesis was rather subtle and indirect in that listeners were expected to shift their boundaries between /d/ and /n/ because of changes in their perception of nasal resonances in coarticulated vowels. Secondly, their subjects were called upon to make a judgment with no linguistic relevance. Vowels as such are not differentiated in English by the feature of nasality. The speakers of American English employed by House and Stevens were asked to judge vowels for nasal quality, much as speech therapists do as part of their auditory assessment of the success of surgical repair of a cleft palate. Thirdly, they used an unnaturally large amount of nasal coupling.

We plan to extend our research on the impedance-matching hypothesis to vowels, but, unlike House and Stevens, we intend to experiment on vowels from a language in which the oral-nasal contrast is phonologically relevant. The experiment which we wish to carry out will require that the chosen language possess pairs of vowels ranging from high to low. This excludes, for example, French in which distinctive nasality is restricted to two vowel heights. A suitable language, we believe, is Hindi. We are now preparing the groundwork for a study based on that language as a sequel to the present paper.

## REFERENCES

Bell-Berti, F., Baer, T. & Niimi, S. Coarticulatory effects of vowel quality on velar function. *Journal of the Acoustical Society of America*, 1978, 63, S33 (A).

Björk, L. Velopharyngeal function in connected speech. *Acta Radiologica Supplementum*, 1961, 202.

Bloomer, H. Observations on palato-pharyngeal movements in speech and deglutition. *Journal of Speech and Hearing Disorders*, 1953, 18, 230-246.

Fant, G. *Acoustic Theory of Speech Production.* 's-Gravenhage: Mouton, 1960.

Finney, D. J. *Probit Analysis*. Cambridge: Cambridge University Press, 1971.

Fujimura, O. Analysis of nasal consonants. *Journal of the Acoustical Society of America*, 1962, 34, 1865-1875.

Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. New York, Wiley, 1966.

Harrington, R. A. A study of the mechanism of velopharyngeal closure. *Journal of Speech Disorders*, 1944, 9, 325-345.

House, A. S. & Stevens, K. N. Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, 1956, 21, 218-232.

Isshiki, N., Honjow, I. & Morimoto, M. Effects of velopharyngeal incompetence upon speech. *Cleft Palate Journal*, 1968, 5, 297-310.

Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, 54, 358-368.

McDonald, E. T. & Baker, H. K. Cleft palate speech: An integration of research and clinical observation. *Journal of Speech and Hearing Disorders*, 1951, 16, 9-20.

Mermelstein, P. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 1973, 53, 1070-1082.

Miller, J. L. & Eimas, P. D. Studies on the perception of place and manner of articulation: A comparison of the labial-alveolar and nasal-stop distinctions. *Journal of the Acoustical Society of America*, 1977, 61, 835-845.

Nusbaum, E. A., Foley, L. & Wells, C. Experimental studies of the firmness of velar-pharyngeal occlusion during the production of English vowels. *Speech Monographs*, 1935, 2, 71-80.

Nylén, B. O. Cleft palate and speech. *Acta Radiologica Supplementum*, 1961, 203.

Passavant, G. *Ueber die Verschliessung des Schlundes beim Sprechen*. Frankfurt a.M.: J. D. Sauerlander, 1863.

Rubin, P. & Baer, T. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 1978, 63, S45 (A).

Rubin, P., Baer, T. & Mermelstein, P. An articulatory synthesizer for perceptual research. *Haskins Laboratories Status Report on Speech Research*, 1979, *SR-57*, this issue.

Warren, D. Nasal emission of air and velopharyngeal function. *Cleft Palate Journal*, 1967, *4*, 148-156.

## FOOTNOTES

[1] The acoustic analyses of nasal consonants, reported by Fant (1960) and Fujimura (1962), have revealed a number of salient features. The first formant typically occurs around 300 Hz and is well separated from the upper formant structure, due to the combined effects of the pharyngeal and nasal cavities determining the fundamental resonance. The increased length of the direct acoustic transmission path causes a reduction in the concomitant average formant spacing and, therefore, a high density of formants in the middle-frequency range. The higher rate of energy loss in the nasal cavity causes greater damping of the resonances with a characteristic increase in the widths of some of the formant peaks; in the case of /n/, the relatively high absorption of sound energy in the termination of the oral cavity shunt causes greater damping of the antiresonance and, consequently, an increase in its bandwidth.

[2] A modification of this specification, termed "partial coarticulation," was used in a pilot study of oral-nasal category boundary movement and in Experiment 1 reported in this paper. The modification involved closing the velum at the end of the tongue movement when it had reached its position for the steady-state vowel. The boundary movement data obtained with these utterances were essentially the same as, although somewhat less robust than, the results obtained with "fully coarticulated" vowels (i.e., when the velum remained lowered throughout the vowel).

[3] As Green and Swets (1966) have pointed out, if one is concerned with sensory events at the level at which they are referred to the perceiver's internal decision space, there are both theoretical and practical reasons for assuming that the underlying distributions of these events are Gaussian. Thus, given a unidimensional stimulus continuum that crosses a decision boundary between two categories A and B on an appropriate psychological scale, the frequency (f(A) and f(B)) of the perceiver's A and B responses to a stimulus at z on the continuum will be governed by the relation:

$$f(B)=K\int_{-\infty}^{z}\exp(-x^{2}/2\sigma^{2})\,dx \qquad f(A)=1-f(B)$$

where K is a constant and $\sigma$ is the standard deviation of the distribution of stimulus plus noise. Since Probit analysis (Finney, 1971) is a procedure for fitting a normal Gaussian integral to sigmoidally distributed data according to a maximum likelihood criterion and since we wished to apply the Probit method to the analysis of our oral-nasal response data, we sought a transformation of the synthesizer's velar port area parameter that would map the output signal distribution into our listeners' *perceptual* space in such a way that their responses at the category boundaries would follow the Gaussian integral.

[4] In nature, the velopharyngeal port is not circular. Björk's (1961) tomographic and cineradiographic study concludes that the velopharyngeal port area is a linear function of the port's sagittal minor axis and that the constant of proportionality is 10 mm. Hence, the anatomical structure of the port is more nearly rectangular. However, for the purpose of computing the degree of oral-nasal coupling, only the total area of the port need be considered and a circular approximation to the velopharyngeal port introduces no significant error.

[5] The evidence was indirect because it rested on the implicit assumption that the relative $F_1$ amplitude and the impedance ratio at the $F_1$ peak frequency are directly related. In practice, this relationship depends critically on the damping characteristics of the nasal tract and becomes less direct as the damping factor is reduced and the opportunity for nasal resonance is enhanced. The nasal tract of House and Stevens' articulatory analogue appears to have been heavily damped and, therefore, they could assume a direct relationship. Since the nasal resonance characteristics of different speakers appear to differ widely and since there are no definitive acoustic data on the subject of nasal damping, it is difficult to make any strong claims for behavioral accuracy from the nasal section of any articulatory model. This is a point made by House and Stevens in their paper and one to which we must also subscribe.

[6] In view of the fact that in moving from an oral to a nasal output, changes can be observed in the structure of many parts of the audible spectrum, it could be argued that any derived impedance ratio purporting to reflect the data available to perceptual mechanisms, should be calculated in such a way as to embrace all the changes rather than those occurring only in the region of $F_1$. Hence, we examined impedances that were derived from the impedance ratio function weighted a) by the magnitude of the first nasal zero; b) by the whole audible spectrum (0-4.9 kHz) and c) by an oral-nasal difference spectrum covering the 0-4.9 kHz range. However, our examination of these other bases for deriving the "perceived" impedance ratio (with the exception of the whole audible spectrum approach) arrived at results which, while differing in minor detail from those in Figure 7, predicted the same basic relationship between the boundary for the /a/ environment and the boundaries for the vowel environments /ʌ/ and /i/. The whole spectrum weighting method led to an ordering of the impedance functions that failed to agree with observation and it was not pursued further.

[7] Had we the time and resources to train thoroughly a small cadre of listeners capable of repeatedly delivering consistent judgments, there can be no question that the experiment could have been completed successfully. However, the very difficulty we experienced in performing the experiment with readily available native English speakers demonstrated to us that the effect at issue was of sufficiently small magnitude in relation to the more easily observed shifts in oral-nasal boundary that we could safely make our choice on an arbitrary basis.

# COARTICULATION AND THEORIES OF EXTRINSIC TIMING*

Carol A. Fowler+

Abstract. Current accounts of coarticulation belong to a single class of theory, here called extrinsic timing theories of speech production. The accounts all assume that the dimension of time is excluded from the specification of a phonological segment in the articulatory plan for an utterance, and all of them fail to explain or predict the coarticulatory patterns of speech. Here I suggest that some of the failings are endemic to the class of extrinsic timing theories, and that a more adequate account must derive from an intrinsic timing theory. The essential characteristics of an intrinsic timing theory are described.

In recent years, several articles addressing the phenomenon of coarticulation as a theoretical issue have appeared in the Journal of Phonetics (Daniloff & Hammarberg, 1973; Hammarberg, 1976; Kent & Minifie, 1977). In the most recent of these, after critically reviewing most of the extant theoretical accounts of coarticulation, Kent and Minifie conclude that none of the reviewed accounts explains coarticulatory patterns adequately. An aim of the present paper is to suggest why current theoretical accounts of coarticulation fail to explain, or indeed adequately to predict, coarticulatory patterns. A second aim is to sketch the form that an adequate account might take. I suggest that our current accounts are all instances of a single class of theory, and that the assumptions about a talker's control over timing in speech that characterize this class preclude its members from providing an adequate account of coarticulation. More precisely, the extant theoretical accounts of coarticulation are instances of theories of extrinsic timing control. That is, they exclude timing from representation in the talker's articulatory plan for his utterance.[1] Instead, they propose that an utterance is given coherence in time only by its actualization.

I propose that some of the inadequacies of our current attempts to explain coarticulation are endemic to this class of extrinsic timing theories, and therefore that a satisfactory account must derive from an intrinsic timing perspective.

*Accepted for publication by Journal of Phonetics, (in press).

+Also at Dartmouth College, Hanover, New Hampshire.

Acknowledgment: I thank Michael Studdert-Kennedy for criticizing an earlier draft of the paper. This work was supported by NIH Grant NS 13617 to Haskins Laboratories.

My procedure will be first to describe the essential, defining properties of extrinsic timing theories. To provide something concrete to work with, I will describe a prototypical theory in this class, drawing in large part on the view of Daniloff and Hammarberg (1973) and of Hammarberg (1976). It is true that current theories of speech production or of coarticulation are substantially different one from the other, particularly with reference to the hypothesized unit of production (for example, the articulatory syllable of Kozhevnikov & Chistovich, 1965, the spatial target of MacNeilage, 1970, and the feature bundles of Daniloff & Hammarberg, 1973). Despite this variety, however, all share the assumptions of the extrinsic timing view. Since these assumptions are the concern here, it seems fair and simpler to focus where possible on a single exemplar.

Having characterized the class of extrinsic timing theories, I will specify why I think no instance can provide an adequate account of coarticulation (or of any of the other manifestations of timing control in an utterance). Kent and Minifie (1977) have catalogued many of the observations that they take to disconfirm the various extrinsic timing theories. The reader who wishes to know the grounds on which the extrinsic timing theories are individually weakened or disconfirmed is referred to that article. Here I will focus on the reasons why (I allege) the theories fail as a class.

In the concluding section of the paper I will characterize the essential properties of a theory of intrinsic timing. The characterization is not intended to constitute an explanation in detail of coarticulation but primarily to specify the theoretical perspective from which, perhaps, an adequate account may derive.

## A PROTOTYPICAL ACCOUNT OF COARTICULATION IN THE EXTRINSIC TIMING FRAMEWORK

An intuitive concept of "segment" underlies our recognition that there is a phenomenon of coarticulation requiring explanation. In a sense, this is unfortunate because the problem of discovering the acoustic or articulatory correlates of our intuitive concept has been recalcitrant. Its recalcitrance stems from an evident incompatibility between the essential properties of the concept and its manifestations in speech (cf. Studdert-Kennedy, 1977). When we perceive speech, we hear a succession of discrete segments. The segments as perceived are discrete or separate in two senses: The successive sounds are both qualitatively and temporally distinct. But when we observe a talker's articulations, or when we look at an acoustic record of them, at most only the first distinction is preserved. One can see a talker producing different kinds of sounds: Some of them occlude the vocal tract locally, some practically obstruct the passage of air, and others merely alter the global shape of the tract. Likewise, one can see the consequently different kinds of acoustic signal. What is not represented in the articulatory and acoustic records of an utterance is temporal discreteness. The different kinds of gestures go on simultaneously, and thus there are no borders perpendicular to the time axis in an articulatory or acoustic record to separate one segment from another.

For Hammarberg (1976), there is only one conclusion:

> Segments cannot be objectively observed to exist in the speech signal nor in the flow of articulatory movements. There are no invariant physical cues of segmentalness. There is no extra-human, i.e. nonsubjective, way of analyzing the speech continuum into discrete parts corresponding to the notion of a segment... What all this adds up to is, that the concept of segment is brought to bear _a priori_ on the study of the physical-physiological aspects of language. (p. 355)

And again:

> It should be perfectly obvious by now that segments do not exist outside the human mind... All indications are that the segment is internally generated, the creature of some of kind of perceptual-cognitive process. (p. 355)

For Hammarberg, then, the upshot is this. The mind has sets of concepts of phonological segments that it _imposes_ on an acoustic signal in the course of perceiving it. However, the segments are not given in the acoustic signal nor in the articulatory gestures responsible for it; thus it takes a human mind to interpret an acoustic speech signal.[2]

Typically, a perceiver of speech hears what the talker intends him to. Thus a statement of the talker's intended utterance, no less than a statement of the hearer's percept, must invoke the concept of segment. Evidently the talker starts with segments, but loses them somehow either in his articulatory plan for his utterance or in its actualization. The charge of a theory of coarticulation is to explain how the mental concepts of discrete phonological segments get translated into a continuous overlapping production of articulatory gestures.

For Daniloff and Hammarberg (1973; see also Hammarberg, 1976), the explanation posits an abstract description of the talker's intended utterance. The plan is a left-to-right array of discrete phonological segments--that is, of the abstract segments that perceivers hear, but that are not given to them in the acoustic signal. Daniloff and Hammarberg call them canonical forms. "They are invariant, ideal, uncoarticulated target forms," each containing all and only that which is essential to their particular identity. Canonical forms never appear in an utterance, but they can be estimated. The best approximation occurs "when a segment is produced in isolation in a sustained manner, or when the sound is produced in a context assumed to be minimally coarticulatory" (Daniloff & Hammarberg, 1973, p. 241).

The dimensions of description of the phonological segments are assumed by Daniloff and Hammarberg to be features (but a compatible proposal is that they are spatial targets; see MacNeilage, 1970). Thus the plan for an utterance at an early stage is a left-to-right array of feature bundles. If the plan were to be executed at this stage, the abrupt changes in articulatory specification that would occur as the plan executor moved from feature bundle to feature bundle would cause transitional sounds to occur between the realizations of sounds that the speaker intended to be immediately successive. To avoid that, features are "spread" in the plan from one bundle to its neighbors so that adjacent sounds and gestures are accommodated one to the other (cf. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In consequence, articulatory transitions between adjacent segments are smoothed, but at the expense of the canonical forms.

It is worth observing that on this view, coarticulation is not coproduction of segments--that is, it is not the overlapping production of separate ideal segments. Rather it is an adjustment of an ideal segment to its context. (Cf. Hammarberg, 1976: "Coarticulation is, then, to be regarded as a process whereby the properties of a segment are altered due to the influences exerted on it by neighboring segments." p. 576)

The foregoing account of coarticulation differs from other accounts both in its degree of precision and in its proposed units of production. But it shares with other accounts several assumptions or claims that identify it as a member of the class of extrinsic timing theories. They are as follows:

1. The essential properties of a segment as it is <u>known</u> to a language user are timeless.

2. Segments in a planned sequence are discrete in the sense that, (abstractly stated), their boundaries are straight lines perpendicular to the time axis, so that the terminus of one segment is the beginning of the next segment. (Feature spreading adjusts the specifications within a pair of boundaries but does not make the segments continuous. Thus, ideally, segments occupy nonoverlapping time slots even after feature spreading.)

   These first two assumptions exclude the dimension of time from having an essential role either in defining the phonological units themselves or their relations in a planned utterance.

3. The plan for an utterance is distinct from its executor. (Concomitantly, the spatial coordinates of an utterance are specified independently of its temporal coordinates. That is, the feature bundles or the spatial targets in the plan specify the successive spatial coordinates of the utterance while the executor or articulatory mechanism actualizes its temporal coordinates.)

None of these assumptions is defended by Daniloff and Hammarberg, or by Hammarberg, despite their unusual efforts to make explicit their assumptions and reasoning. Indeed none has been defended to my knowledge by any theorist of coarticulation or speech timing. However, the claims are addressed by Lashley (1951, pp. 506-528) in his classic paper on serial ordering. Since I will argue that they are false, a digression to examine and evaluate his arguments is warranted.

## LASHLEY: THE ROOTS OF EXTRINSIC TIMING THEORIES

Many investigators and theorists recognize the relevance of the classic serial ordering issue to that of coarticulation. In that connection, it is not surprising that Lashley's well-known paper "The Problem of Serial Order in Behavior" (1951) is frequently cited when coarticulation is given theoretical treatment (e.g., Hammarberg, 1976; Kent & Moll, 1975; MacKay, 1970; Wickelgren, 1969). Nor is it surprising that his views have their counterparts in accounts of coarticulation.

Lashley presents the extrinsic timing view in the following passage:

> Since memory traces are, we believe, in large part static and persist simultaneously, it must be assumed that they are spatially differentiated. Nonetheless, reproductive memory appears almost invariably as temporal sequence, either as a succession of words or of acts... The translation from the spatial distribution of memory to temporal sequence seems to be a fundamental problem of serial order... There are indications that one neural system may be held in [a] state of partial excitation while it is scanned by another... The scanning of the spatial arrangement seems definitely to determine, in such cases, the order of procedure. (pp. 521-522)

I understand Lashley to make the following argument. Some acts (in particular those like speaking in which there is a more-than-additive relationship between the sense of the act as a whole and that of its components taken separately) presuppose "mental plans." A plan must represent concurrently the act-components that will occur in succession when executed. Since the representations of the various act-components (Lashley's memory traces) persist concurrently and over some period of time, the plan is time-invariant. Thus the dimensions along which the representations may encode information about their referent act-components must be the three spatial dimensions at most. Another consequence of the plan's time invariance is that the memory traces representing successive act-components must be spatially differentiated. Only on execution of the plan are they serially ordered in time and given temporal coherence. Lashley's conclusion that time cannot be represented in the plan, then, derives from the evident fact that it cannot be given literal representation because of the necessarily static nature of the plan's components.

This is the essence of the extrinsic timing view and it survives intact in our accounts of coarticulation. But Lashley's argument, if I have characterized it fairly, is specious. He seems to conclude that because a memory trace is static--and thus at most three-dimensional--it can only encode information about those three dimensions. But of course this is not a limitation. The written sentence: "John is eating an apple" is a representation in two dimensions of a four dimensional event. The information conveyed by a representation is not constrained by the dimensionality of the latter. Thus although information conveyed by a spatial array of static memory traces could, by coincidence, be information about successive, static parts of an act, it need not be.

This means only that we are not bound to devise extrinsic timing theories--at least not on the grounds specified by Lashley. It does not imply that extrinsic timing views must be incorrect.

Let us turn now to what I see as the failings of the extrinsic timing theories. I will consider each of the assumptions of these theories described earlier.

## CRITICISMS OF EXTRINSIC TIMING THEORIES

### Assumptions 1 and 2: Sequences as Timeless; Intersegmental Boundaries as Abutting Straight Lines Perpendicular to the Time Axis

I will consider together the first two assumptions listed above--that segments are timeless and that their boundaries are abutting straight lines oriented perpendicularly to the time axis. The two assumptions are distinct in the extrinsic timing theories, but the reasons why they are incorrect overlap in large part.

In general I will argue that the concepts of a segment as, ideally, a three-dimensional layer in a four-dimensional event of talking, and of talking itself as a succession of those layers, must be incorrect. Furthermore, their invalidity is sufficient to preclude their constituting a model for a talker of his own utterance. That is, they characterize neither an utterance nor a talker's plan for his utterance, nor the "canonical forms" that underlie them.

First consider some counter-evidence to the proposal that a segment is timeless so that it can be specified in a plan as the (planned) convergence at an instant in time of a bundle of features or as a spatial target. Lisker (1972, pp. 2387-2418) provides one reason why this point of view must be incorrect. He notes that some languages have phonological segments, such as pre-nasalized stops or occluded nasals, which cannot be characterized as a complex of simultaneous features. These segments involve a sequence of necessarily non-overlapping states of the velum.

This observation may have its counterpart in some sounds of English as well. Bell-Berti and Harris (personal communication) found the onset of coarticulatory lip rounding for /u/ to precede the acoustically defined onset of /u/ by a fixed amount of time. That is, in their data, lip rounding was not time-locked to any of the segments preceding /u/ in the plan. Rather, rounding was time-locked to the remaining gestures for /u/. Hence, it would seem, /u/ is a four-dimensional segment, which is co-produced with its neighboring segments.

In a similar vein, Kent, Carney and Severeid (1974) describe the production of the first nasal /n/ in "intend." Velar closing in anticipation of the following non-nasal consonant is synchronized to the tongue gestures that produce the /n/. By the time the alveolar tongue constriction for /n/ is attained, the velum is already half way to its closed state. Whatever it is that specifies the coherence of those gestures that correspond to a given segment, it cannot be synchrony, because if it were, the /n/ in "intend" would be a non-nasal.

Still focusing on the individual segment, the work of Abramson and Lisker (e.g., see Abramson, Note 1) indicates that voiced and voiceless segments are distinguished on the articulatorily simple (but acoustically complex) dimension of laryngeal timing. Finally, Lisker (1972, pp. 2387-2418) notes that the phonemes /t/ and /č/ are distinguished one from the other by their relative rates of release of the alveolar constriction. /b/ and /w/, and /d/ and /y/ are similarly distinguished (Liberman, Delattre, Gerstman, & Cooper, 1956). But in order to incorporate "rate of release" as a property of a phonological segment, the phoneme's specification as a spatial target or as the convergence at a point in time of a set of features must be abandoned.

The concept of a plan as an array of feature bundles may also be evaluated on existing evidence. Recall that feature bundles in a plan are context-adjusted. The articulatory mechanism (the plan executor) successively reads out the adjacent feature bundles in the plan. Thus, the features in a bundle are triggered at once, rather than over time.

A prediction may be derived from this characterization that allows it to be tested. When the articulatory mechanism executes a context-adjusted feature bundle, the component features will be realized as articulatory activity with mutually somewhat different latencies. Each latency will depend on the articulator implicated by the feature and that articulator's current activities. It may be possible to identify, either in the EMG (electromyographic) signals for the relevant muscles or in the movement records of an utterance, the relative latencies of the different features that were simultaneously executed by the articulatory mechanism. Barring variation in carry-over coarticulation, these relative latencies should be invariant over any change in context. The latencies should reflect the invariant time that it takes each feature to be actualized when a bundle is executed. In short, the evidence should never indicate that the different members of a bundle are triggered over time rather than simultaneously.

In accordance with the feature-bundles view is evidence reported by Kent and his colleagues (Kent & Netsell, 1971; Kent, Carney, & Severeid, 1974; Kent & Moll, 1975) that articulatory gestures are synchronous or time-locked over very short intervals. If this occurs universally, it supports the view of the plan as an array of feature bundles.

It does not occur universally, however, as some EMG and cineradiographic data show. The decision as to when a particular gesture is to be initiated sometimes seems to be made with reference to when it has to be initiated in order to get the job done, rather than with reference to the onsets of other gestures. For example, Bell-Berti (1973) shows that activity of the levator palatini (which is necessary to raise the velum after a nasal) in /fVCmVp/ begins earlier within the nasal in anticipation of the vowels /i/ and /u/ than in anticipation of /a/. Bell-Berti suggests that this occurs, not for any mechanical reason, but because nasal coupling is less likely at a given degree of opening during the production of a low vowel such as /a/ than during the production of high vowels /i/ and /u/. It is as if the system delays velar closure until it has to initiate it. Some cineradiographic data bear out the EMG evidence that individual articulatory gestures need not be time-locked. Borden and Gay (1975) describe tongue, jaw and lip movements during the productions of /spapə/, /stapə/ and /skapə/. Although time-locking was evident for one of three speakers producing /spapə/, it was not for the other two speakers, nor generally for any of the speakers producing /stapə/ and /skapə/.

One might seek to revise, rather than to reject the feature-bundle notion by supposing that at a finer grain of description, features are arrayed sequentially as well as in bundles in the plan (see Kent, Carney, & Severeid, 1974). I think that there is a more satisfactory solution, however, which closes the separation between the properties of canonical forms and those of their actualizations.

In the discussion that follows I will make the following arguments:

1. Any proposal in which distinctions are posited between a canonical form and its actualization (especially distinctions as difficult to bridge as the discrete/continuous, timeless/four-dimensional differences proposed in extrinsic timing theories) is an argument of last resort. This is so in part because it leads one to ask why a conceptual category should arise in evolution or ontogeny that bears so little resemblance to the actualization with which it co-evolves or co-develops. But it is also unattractive because it vastly complicates the process of getting from canonical form to gesture and back again in perception. (As Hammarberg notes, it introduces the mind/body problem. It does so because it demands translations across phases of matter--from psychological to physical and from physical to psychological.) I argue that we need not yet give up on our search for the correlates of our intuitive concept of segment in a speech utterance or in its acoustic product.

2. The separation between canonical form and actualization leads necessarily to the notion of Hammarberg's that phonological categories are brought to bear *a priori* (by a phonetician, and presumably also by a naive listener) on an acoustic signal. To the extent that this implies insufficient acoustic support for the conceptual categories that we perceive, the claim is untenable and must be gotten around.

## Canonical Forms

A speech production theory must be more highly valued, other things being equal, if it posits no difference between the properties of canonical forms of phonological segments (that is of segments as they are known) and those of segments described in an articulatory plan or realized in a vocal tract. That is so because any difference requires both a "translation theory" (cf. Fowler, Rubin, Remez, & Turvey, in press)--that is an explanation of how the translation is effected from known segment to produced segment--and a rationalization for the evolution of concepts to be communicated that cannot be nondestructively actualized (see below). I assume that theories treat canonical forms as three dimensional and as ordinally separate from their neighbors (as of course they are not in speech itself) in part because theorists believed themselves backed into those decisions (due to the reasoning made explicit by Lashley and described earlier) and in part because they considered linguistic theory to verify these as properties of abstract linguistic segments (cf. MacNeilage & Ladefoged, 1976, pp. 75-120). Since sequences of segments cannot be static when realized in a vocal tract, actual segments cannot share this property with segments as speaker/hearers know them. Furthermore, certain cognitive demands of speech perception seem to require coarticulation (Liberman & Studdert-Kennedy, 1978), and coarticulation according to some interpretations precludes actualized segments being discrete and context-free.

However, elsewhere (Fowler et al., in press), I have suggested that segments as formal linguistic entities do not in fact have the properties 'static' or 'discrete.' Formal linguistic theory seeks only to characterize segments as they participate in an abstract linguistic system. Because it concerns itself only with an abstract formal system, the theory is unconcerned with the way in which these formal properties are known, produced or perceived by speaker/hearers. Thus the theory assigns featural values to segments only on dimensions of description that distinguish one segment from its class members. The dimensions 'static'/'dynamic', and 'discrete'/'continuous' are irrelevant here because they have to do with the way in which segments are realized in some medium and that is not of concern in a formal linguistic description.

Whether or not this strategy of formal linguistic theory is useful is not of concern here. What is important is that linguistic theory does not in fact assign the values static and discrete to linguistic segments. Hence, the theorist of speech production is free to assign whatever values are most useful to him on the dimensions static/dynamic and discrete/continuous. Known and produced segments may <u>both</u> be four dimensional, (i.e., [+dynamic]). The assignment of this value to linguistic segments does no violence whatever to their other canonical properties as assigned by linguistic theory. Similarly, instead of treating coarticulation as an adjustment of the canonical properties of a segment in acquiescence to its neighbors, it may be viewed as the <u>overlapping</u> production of successive, continuous, four-dimensional segments. Thus feature spreading may be apparent, but not actual.

If segments are coproduced, and if they are not temporally discrete even in intent, then why do our introspections reveal them to be distinct? Our introspections may yield an impression of discreteness for two reasons. First and foremost, the phonological segments that are coproduced--primarily consonants and vowels--are different <u>kinds</u> of segment. The one is a rapid and local obstruction of the vocal tract, while the other is a relatively slow global change in the shape of the vocal tract. These different kinds of gesture generate different kinds of acoustic signal; crudely, clear cases of consonants provide acoustic evidence of vocal-tract constriction, while vowels reflect the formant structure of an open tract. Just as we perceive the separateness of any two acoustic events of different kinds that overlap in time (say of a singer's voice on a record and of her musical accompaniment), we can (putatively) detect the more subtle separateness of two kinds of spoken segments that overlap in time. The second reason for our perception of spoken segments as discrete may be that no two sounds are strictly concurrent. For example, a vowel and a consonant may be coproduced, as in the data on VCV production reported by Ohman (1966); but the onset of the first vowel precedes the consonantal constriction, and the second vowel persists after the local occlusion for the consonant has been released. To repeat, then, the suggestion is that segments are not temporally discrete, but rather are qualitatively separate events, and that this separateness accounts for our impression that segments in an utterance are mutually distinct.

If we accept that segments are essentially four dimensional and are coproduced, an advantage is that we need not view the velum-raising data of Kent et al. (1974), for example, as a paradoxical case of feature spreading. Instead, the data may indicate that a nasal segment is four dimensional and that its properties are revealed over time. Its coherence is specified not by the convergence of its featural actualizations at a point in time, but by the continuity in time of the production of a segment of a particular kind. Likewise, anticipatory lip protrusion for a rounded vowel may occur, not because the feature [+rounding] has been spread from the rounded vowel to earlier segments, but because rounded vowels are entities of which time is an inherent dimension--they are produced in time.

## A Priori Categories

Let us next consider Hammarberg's conclusion that since canonical forms are not actualized unaltered in an utterance, the concept of segment must be brought to bear a priori on the physical acoustic signal.

The suggestion that phonological segments are subjective categories leaves unanswered several puzzling questions. For example, how could the a priori concept of segment ever have arisen in evolution (i.e., if the concept cannot be learned by a child from his experiences with the acoustic signal, how could it ever have been acquired by a species based on its experience with acoustic signals)? Second, why should these fictitious concepts ever have arisen in evolution? Evolution, like ontogeny, is a process by which an organism maintains or enhances its compatibility with the properties of the world. It would appear highly disadvantageous for an organism to seek to impose on the world properties that the world does not have. Finally, if an acoustic signal offers insufficient support for the concept of segment, why do listeners so reliably perceive a talker's intended message?

It is surely more plausible to suppose that the concept of segment has material support. Its essential properties are manifest in the acoustic signal, although it may take a human perceptual system to detect that aggregate of properties as a significant collective. Scientists have not discovered those properties in the acoustic signal, but the reason they have not may be that they have looked for evidence of the wrong kind. They have looked for temporal discreteness when they should have looked for qualitative separateness among temporally overlapping events. And they have sought to discover abutting edges of segments perpendicular to the time axis when, perhaps, no such things are to be found.

Some theorists (e.g., Gibson, 1977, pp. 67-82; Polanyi, 1958) take issue with the dichotomy between subjective and objective events to which Hammarberg makes reference in his discussion of a priori concepts of segments. The dichotomy as applied to the concept of phonological segment suggests that segments as perceived and known are either objectively instantiated--that is, their essential properties are manifest in the acoustic signal and thus provide support for the perceiver's cognitions--or their properties have no acoustic support and the concept of segment is in Hammarberg's (1976) words "internally generated, the creatures of some kind of perceptual-cognitive process" (p. 355). But some theorists prefer a third alternative. The third alternative is that percepts and concepts are neither subjective nor objective; instead their material support is available in the world, but is only detected by a specially attuned organism. Consider again the example of footwear as discussed in footnote 2. Clearly the aggregate of properties--foot-shaped and -sized, protection for the foot, etc.--are available in the light to an eye when that eye focuses on an instance of footwear. Thus the essential properties of the concept are "objectively" manifest. But it takes a wearer of shoes to detect that particular aggregate of properties as a significant collective.

It is implausible to suppose that the concept of phonological segment is wholly subjective--that is, that it has no acoustic support. It seems more reasonable to suppose that 'phonological segment' is a concept like that of footwear--it is neither subjective nor objective, but something else that spans the dichotomy.

## Assumption 3: The Plan as Distinct from its Executor

The recent literature provides both counter-evidence (Monsell & Sternberg, in press) and counter-arguments (Neisser, 1976; Fowler, 1977) to the view that plans are executed by an extrinsic articulatory mechanism. The counter-evidence derives from an experimental paradigm in which subjects are asked to produce a well-learned or well-known utterance as quickly as possible following a signal. Subjects know in advance of the signal what utterance they are to produce; the signal simply indicates when they are to begin talking. Monsell and Sternberg observe a positive linear relationship between latency to begin producing the utterance following the signal and a measure of utterance length (the number of stressed syllables in the utterance). The result holds, even though the utterance is known to the subject before the signal to respond, and thus, even though he evidently has ample time to prepare for it. It also holds when the sequence to be produced is very familiar to the subject (e.g., some portions of: "one, two, three, four, five": or "Monday, Monday, Monday, Monday, Monday"). Monsell and Sternberg suggest that the reason why subjects fail to "take advantage" of the time before the presentation of the signal to devise a plan for their utterance is that plans are self-executing. When a plan for an utterance is devised, it necessarily runs off. The intrinsic timing view that I will sketch out below provides a rationalization for this result.

In addition to this limited counter-evidence, there are strong grounds for questioning the proposal that plans are separate from their executors. The reasons concern the obligations of a proposed model of coarticulation. A model of speech production, even if devised to explain only coarticulation, must still accommodate (or must enable elaboration that will allow it to accommodate) other manifestations of timing control. That is to say, we have sufficient grounds for eliminating a theory of coarticulation if the model of speech production that it promotes could never be made to generate well-known coarser-grained timing phenomena, such as rate effects, stress-timing and initial and final lengthening, in a natural or plausible way.

In fact, extrinsic timing theories do not adequately handle either rate effects or stress-timing. Consider rate first. Vowels produced at a fast rate are substantially shorter in duration than are their more leisurely counterparts. Their reduction in duration seems to be accomplished by a reduction in their "target" spatial coordinates (e.g., Harris, Note 2). That is, the articulatory movements towards the canonical target (as estimated from the vowel produced in isolation) are less extensive for vowels spoken at a fast rate than for vowels produced at a comfortable rate. Apparently this is because they are produced w th less effort or muscular force than slower vowels. See Gay and Ushijima, 1974. A typical concomitant is that their midvowel formant values are centralized (Lindblom, 1963).

This means of increasing speaking rate for vowels does not, and could not, have its parallel in respect to consonants. Rapidly spoken consonants are slightly shorter in duration than are leisurely consonants, and they are evidently produced by *increasing* effort or force to the relevant muscles (Gay & Hirose, 1973; Gay & Ushijima, 1974; Gay, Ushijima, Hirose, & Cooper, 1974). Indeed, consonants could not be produced rapidly by decreasing muscle force, because their essential articulatory properties include obstructing or totally occluding the passage of air from the lungs. If the muscle forces that effect that consonantal obstruction were diminished, a different class of segment would be produced.

This evident difference between vowels and consonants must be mimicked by a model of speech production and must be rationalized by its accompanying theory. However, it is not mimicked in a natural way by a model in which feature bundles for consonants and vowels are laid out in left to right adjacency in a plan and are executed by a separate articulatory device. That device must approximately alternate, first executing a vowel at a slow rate (or equivalently first assigning a reduced-from-normal amount of muscular effort), and then reading out a consonant at a fast rate (or assigning it super-normal muscle force). But this dual strategy seems only to complicate the talker's task. Why does he not simply increase muscle force generally and produce all sounds according to the consonant strategy? Although the extrinsic timing device can generate these rate effects, it does so only in a post-hoc way that fails to rationalize a dual strategy.

The conclusion is similar in regard to stress-timing. There is now available fairly substantial evidence that English speakers regulate the intervals between stressed vowels in an utterance (see Fowler, 1977; also see Sternberg, Monsell, Knoll, & Wright, 1978), and there are grounds for supposing that an extrinsic timing model cannot generate those intervals. The evidence of Morton, Marcus, and Frankish (1976), of Rapp (Note 3) and of Allen (1972) suggests that the controlled intervals lie between the *centers* (however defined) rather than the *edges* of any linguistic units, including the phonological segments. If this is so, timing constraints can not be between feature bundles. However, even if an extrinsic timing theory could be made to generate stress-timed utterances, it can not rationalize the phenomenon. Indeed, like the different rate strategies for vowels and consonants, stress-timing would simply be an added burden for a plan executor.

However, it seems most unlikely that speech production would be easier for a talker if he were not obliged to produce special rate effects or stress-timing: these phenomena must not be treated in a model as if they were merely arbitrary complications. Rather they are, and should be treated, as indicants of the strategy or style of control that talkers adopt when they produce an utterance. An adequate theory of timing control, then, is one that describes a style of control out of which these traces of timing control fall naturally. In the concluding section I will describe the theoretical perspective from which a more adequate account of coarticulation may perhaps be derived.

# COARTICULATION AND INTRINSIC TIMING

## Obligations of an Intrinsic Theory of Coarticulation

The obligations of an intrinsic timing theory of coarticulation are at least these four:

1. The theory must characterize the essential properties of segments as four-dimensional entities.

2. More abstractly, the theory must rationalize the classification of segments into vowels and consonants. This is evidently necessary if the model of coarticulation is to be compatible with the coarse-grained timing phenomena of rate and stress-timing (and if it is to capture our intuitions that consonants and vowels constitute different kinds of entities).

3. The theory must merge the plan and its executor by incorporating time into the plan for an utterance.

4. The theory must rationalize coarticulatory effects, and a model derived from it must be able to generate them.

Below I will characterize a way of conceptualizing speech production that may meet these obligations. The description is of necessity terse. For a more elaborate discussion, see Fowler, 1977, and Fowler et al., in press.

## Coordinative Structures

All activities that we perform are coordinated. (This is true of even our clumsiest actions. Compare them with the maximally uncoordinated convulsion.) In order for acts to be coordinated, the muscles that contribute to their realizations have also to be coordinated. If they were not, then different muscles would compete and unorganized movements would ensue. These functional organizations of muscles are called coordinative structures by Easton (1972). Their governance of acts such as locomotion (Easton, 1972; Grillner, 1975), swallowing, and chewing (Doty, 1968, pp. 1861-1902; Sessle & Hannam, 1975) are well documented.

Significant properties of a coordinative structure are that it generates an equivalence class of movements, that it is nested, and that it frequently is cyclic in nature.

In respect to the first property, the coordinative structure is an organization that spans several muscles and produces activities of a certain *kind*. The organization over the muscles may be described as a mapping. For example, that which regulates movement of the forearm at the elbow under some conditions may be described by the equation for a nonlinear spring: $F = ae^{k(l - l_o)}$ (Fel'dman, 1966). The mapping has parameters, some of which are under the actor's control. In the mapping above, they are k and $l_o$. When those parameters are given different values, different but similar movements ensue. Thus the organization described by the mapping engenders a *family* of acts.

The second property of coordinative structures--that they tend to be nested--is evidenced in the act of walking. During locomotion, small systems of muscles that govern intralimb stepping are nested within (or organized into) a large muscle system whose role is to coordinate the activities of two (or four) limbs (see Easton, 1972). Notably, the "life-span" of the small coordinative structures--that is the duration of time over which the small coordinative structures function--is shorter than that of the superordinate muscle system.

Finally, many coordinative structures, including those involved in walking, chewing, and in respiration, are cyclic. That is, once a muscle system has run through its repertoire of activity (in walking, once a single step has been taken by each limb), the repertoire is reinitiated--that is, the "end" of a cycle reinitiates the sequence.

It is important to recognize the efficiency of this style of organization. Coordinative structures are self-executing organizations: once they have been marshalled, no further organizational intervention is required as they execute their special repertoire of activity. If an act is cylic--like walking, breathing, chewing (and I will suggest vowel production)--acts of indefinite duration may be evoked by just once marshalling the requisite muscle organization.

The reader is referred to Turvey (1977, pp. 211-265) and to Greene (1972, pp. 304-335) for a verification of these properties of organized muscle systems. I will suggest below that a plan as a nesting of coordinative structures meets the obligations listed earlier for a model of intrinsic timing.

## Coordinative Structures in Speech

First I will describe some of the coordinative structures involved in the act of speaking--both as regards its components that are considered to be involuntary and automatic and as regards those components that are considered voluntary. Next, I will suggest how that proposed style of control meets the obligations of a theory of coarticulation.

The respiratory system. Basic reflexes operate both in vegetative breathing and in speech, apparently to regulate the initiation and termination of inspiratory activity (see, for example, Kaplan, 1971). Briefly, the expiratory "center" of the brain stem receives inhibitory innervation from stretch receptors in the alveoli of the lung. The expiratory cells inhibit their inspiratory counterparts and, ceteris paribus, terminate their activity. Chemoreceptors in the vicinity of these brain stem centers that are sensitive, for instance, to the level of $CO_2$ in the blood also regulate the centers' activities. These reflexes operate like control systems in working to regulate the $CO_2$ level of the blood.

When viewed more macroscopically than this, vegetative respiration and speech respiration are quite different. During vegetative breathing, the activity of the inspiratory muscles is in phase with the inspiratory portion of the respiratory cycle (Lenneberg, 1967). Furthermore, except during forced expiration, that phase is typically accomplished passively--that is, by relaxing the muscles of inspiration and by allowing the elastic recoil forces of the lungs to work unaided and unopposed. The expiratory phase occupies about 60% of the cycle.

But during speech the sequence of events is somewhat different (Draper, Ladefoged & Whitteridge, 1959; Lenneberg, 1967). In speech, the activity of the inspiratory muscles is out of phase with the act of inspiration. Their activity extends into the early part of expiration where they act to check the descent of the rib cage that characterizes passive expiration. Immediately following the decline and offset of activity in the inspiratory muscles, the internal intercostals that are muscles of active expiration come into play. If phonation is prolonged, two other muscles that may contribute to expiration, the rectus abdominis and the latissimus dorsi, are also marshalled (Draper et al., 1959).

Two results of this coordinated activity are that the proportion of the respiratory cycle occupied by the expiratory phase is about 0.87 (Lenneberg, 1967) and that subglottal pressure is maintained at a nearly constant level despite the continual decrease in the volume of air in the lungs (Draper et al., 1959; Lieberman, 1967). Lenneberg refers to these coordinated activities of the muscles of inspiration and expiration as "synergisms," a term that others have used as Turvey (1977, pp. 211-265) and Easton (1972) use the term coordinative structure. In the case of this macroscopic coordinative structure, it is a device whose task is to control subglottal pressure, much as the microscopic reflexes regulate stretch and $CO_2$ levels in the blood. The macroscopic coordinative structure is superimposed on the smaller reflexes when an individual chooses to speak.

Notably, in the experiment of Draper et al., similar sequences of muscular events occurred over a range of controlled subglottal pressures. That is, the same coordinative structure may govern an utterance over all amplitudes of production.

The laryngeal system. Evidence that the operation of autonomous laryngeal devices contributes to speech production is sparse. What evidence there is, is provided primarily by Wyke (1967, 1974). In his view:

> The production of speech by human beings is, in essence, similar to a large number of other acts of daily life that the human being fashions unthinkingly out of coordinated variations in the tone of striated muscles (See Wyke, 1959, 1967b). Obvious examples of such automatic acts performed with striated muscles are walking, mastication, swallowing and breathing. In all of these situations, the muscular performance is certainly capable of voluntary initiation, modification and arrest; but its detailed production, in the circumstances of normal everyday life, is continuously adjusted by reflex mechanisms that operate at a subconscious level, and over which we have no voluntary control. Speech is in a similar situation, as far as phonation goes: that is to say, the production of speech although initiated voluntarily, is dependent mechanistically upon the precise subconscious integration of a large number of feed-back (servo-) reflexes which constantly adjust the tone of the large number of muscles involved in the production. (1967: 2-4)

Wyke has identified three servo-systems in the larynx itself. But the role of these systems in natural speech, if any, has not been established.

Receptors in the mucosal membranes of the larynx are sensitive to changes in air pressure. Wyke (1967) suggests that the afferents from these receptors may trigger reflex adjustments in the tone of the laryngeal and respiratory musculature during speech. Evidence of this, summarized in a recent paper (Wyke, 1974), shows that when subglottal pressure increases during phonation, the tone of the vocal fold adductors is increased and that of the abductors is concomitantly decreased. The adaptive result is that the folds resist the "upward ballooning" that would otherwise follow an upward surge in subglottal pressure.

In addition, Wyke and his colleagues (summarized in Wyke, 1967) have identified mechanoreceptors in all of the laryngeal joints (thyro-epiglottic, thyrohyoid, cricoarytenoid, cricothyroid). The receptors are structurally identical with skeletal joint receptors. They respond to displacement of the joint and their effect is to change the tone of intrinsic laryngeal muscles.

Finally, mechanoreceptors in the intrinsic muscles of the larynx respond to stretch and reflexively alter the tone of the intrinsic muscles.

Superimposed on these microscopic reflexes may be more macroscopic coordinative structures that are responsible for establishing the different laryngeal "modes" of phonation (including normal phonation, whispering, falsetto, creaky voice, etc.). These modes of phonation are established by adjusting, for an extended period of time, various properties of the larynx. According to Abercrombie (1967), the kinds of long-term adjustments that languages may prescribe, or in the case of whispering or the falsetto register, that speakers may choose to adopt, are the following:

> The glottis may be entirely in vibration, or only in part and the part that is not in vibration (usually the so-called cartilage glottis) may be firmly closed, or may be sufficiently open to allow air to pass through at that point; two parts of the glottis may be in different modes of vibration simultaneously; the whole larynx may be raised or lowered in the throat; and the parts of the larynx above the glottis may or may not be constricted. For example, what was called above "breathy" phonation is produced by part of the glottis being in vibration while the cartilage glottis is sufficiently open to allow air to pass freely through it... (1967: 100)

The supralaryngeal system. On a microscopic scale, Folkins and Abbs (1975) provide suggestive evidence for a closed-loop, jaw-lip system that operates during speech production. In their experiment, they asked subjects to produce the phrase "a hæ 'pæp again" repeatedly. On about one-quarter of the repetitions, randomly interspersed, the experimenters applied a transient disturbance to the jaw. The disturbances were applied during the course of lip closure for the first /p/ in the phrase, and were such that they prevented the jaw from reaching its usual degree of elevation. Nonetheless, on every repetition, perturbed or not, the subject attained lip closure due to exaggerated upper and lower lip displacements, and slightly exaggerated velocities of lip closing gestures. These results are explained most simply in terms of a low-level jaw/lip control system that is responsible for lip closure.

Several researchers (for instance, Kozhevnikov & Chistovich, 1965; MacNeilage, 1970; Ohala, Hiki, Hubler, & Harshman, 1968) have observed that the velocity of jaw closure for any consonant varies directly with the distance that the jaw has to travel to attain closure. The consequence is a nearly constant duration of the closing gesture independent of its extent. It is likely that the tongue tip behaves similarly as Sussman (1972) points out. Ohman's (1967) data show nearly identical formant transition durations in /idi/ and /ada/ although the tongue tip has to travel farther in the latter case in order to attain closure. The typical interpretation of these findings is that the velocity of movement is under closed-loop control such that it is adjusted to current conditions.

The supraglottal structures as well as the respiratory and laryngeal structures participate in the coordinated movements that are involved in chewing and swallowing. These acts quite clearly are products of nestings of synergies or coordinative structures (see for instance, Doty, 1968, pp. 1861-1902; Sessle & Hannam, 1975). Doty (1968, pp. 1861-1902) has shown that selective destruction of parts of the "swallowing center" of the brain stem leaves intact some of the "subsynergisms" of the act of swallowing while impairing others. He suggests that some of these subsynergisms resemble speech gestures and may be marshalled during speech. In particular he suggests that the movements of the soft palate during speech are "lessened" forms of those that occur during swallowing.

Others (Bosma, 1953; Fawcus, 1969; Shohara, 1936) have made similar suggestions with regard to the synergies involved in the acts of chewing, swallowing and sucking. Shohara (1936) suggests, for instance, that the tongue gestures during the production of /k/ and /g/ are versions of those involved during swallowing. According to Fawcus (1969):

> Most, if not all of these fine movements of the oralpharyngeal structures (in speech) occur during mastication in both non-speaking children and other nonspeaking animals. The harnessing of these basic movements by the CNS is the essential problem in both the normal development of speech and the procedures designed to overcome developmental failure. (p. 558)

But none of these claims, to my knowledge, is based on any hard evidence that speech gestures are the products of synergies or coordinative structures, differently organized, that participate in other oropharyngeal acts. The evidence appears rather to be informal. Some speech gestures resemble those involved in other acts that utilize the same structures.

Stronger evidence for macroscopic coordinative structures in speech derives from the speech production literature itself. One way in which researchers in other domains identify a coordinative structure is by means of its muscular or gestural concomitants. Recall that a coordinative structure is a group of muscles constrained to act as a system. It generates an act whose properties are stereotyped.

The speech literature provides some examples of this.

1. Kent and Netsell (1971) find that the gestures of the tongue body and the lips are synchronized during the production of the word "we" in "we saw you." A figure relating the displacement of the tongue body to that of the lips over different stress patterns of the phrase (we saw you; we saw you; we saw you) shows that the relationship between the two variables is invariant over differences in stress. Kent and Netsell obtain a similar result for the diphthong /ɔi/ in "convoy" and "convoy," and tentatively conclude that "for sounds like /wi/ and /ɔi/ which are characterized by coordinated movements of two articulators, the stress contrast must alter both gestures or neither gesture" (p. 40).

2. Kent, Carney and Severeid (1974) conclude on the basis of their data that in many cases "articulatory movements seem to be programmed as coordinative structures so that movements of the tongue, lips, velum and jaw often occur in highly synchronized patterns" (p. 487). Some of the data of Borden and Gay described earlier also reveal synchrony in the movements of different vocal tract structures.

3. Of greater interest are cases when the gestures are not synchronized, but rather occur in a constrained pattern over time. Notice that evidence of this can be obtained only by comparing similar utterances or by comparing the same utterance produced at different rates. That is, a nonsynchronized pattern can be detected only if it remains invariant in different contexts. The recent data of Bell-Berti and Harris, cited earlier, showing that the onset of lip rounding precedes the measured acoustic onset of /u/ by a relatively fixed interval, regardless of the preceding consonantal context, seem to provide an instance of this.

   Kent, Carney and Severeid provide some limited evidence that the relative timing of gestures of the tongue body, velum and lips tends to be invariant over a change in rate of speaking. For each articulator and for each of two speakers, figures are provided that superimpose the movement tracings at two different rates during the production of "soon the snow began to melt." The rates of speaking were in the ratio 2:1. To facilitate a comparison of the movements' relative timing at the different rates, the investigators compressed the time-scale of the slower movement relative to that of the fast movement. For both speakers, and for all three articulators, the tracings at the two rates of production were nearly identical. Thus the relative timing of the movements of a particular articulator and the timing relationship among articulators remained nearly invariant over a two-fold increase in rate. These findings are interesting, but questionable in that they seem to conflict with other evidence. For instance, Gay and Ushijima (1974) show that the muscle activity for consonants is of greater amplitude and that of vowels of lesser amplitude during rapid than slow speech. Of course, these investigators provide only EMG data, and it is difficult to make inferences from muscle contraction to movement.

## Coordinative Structures that Encompass the Respiratory, Laryngeal and Supralaryngeal Systems

Obviously, unless the respiratory, laryngeal and supralaryngeal systems are in fact separate systems, there is no need to ask how they are coordinated. But the perspective on the action system suggested by Greene (e.g., 1972, pp. 304-335) and by Turvey (1977, pp. 211-265) is one in which coordinative structures are nested, and these three systems seem to represent a natural fractionation of the whole. Some limited evidence that the three systems or subsets of two of them are coordinated during speech is provided by the studies briefly described below.

1. Perkell (1969) suggests that the "intent to lengthen the vocal tract" and the intent to shorten it are actualized by a coordinative structure controlling the lips, jaw, hyoid and larynx. In his words:

   > It follows that there is a physiological as well as an anatomical interaction between the lips, mandible, hyoid bone and larynx which causes these structures and the muscles connecting them to operate as a unit in shortening the vocal tract (to raise formant 1 and lower formant 2) for the three non high vowels /a,æ,ɛ/... the familiar lip rounding function, combined with vocal tract lengthening at the laryngeal end, comprises a physiological mechanism operating in opposition to the vocal-tract shortening function. (pp. 40-41)

2. Similarly, in the experiment of Folkins and Abbs (1975) described earlier, the investigators compared lip movements in the presence or absence of jaw movement impedance. The displacement of the lips was greater when jaw movement was restricted. However, velocity of lip closure did not increase proportionately on those trials. Therefore lip closure was attained 15-25 msec later on the test trials than on the control trials. Folkins and Abbs observed that voicing offset was also delayed on those trials, suggesting to them a coordination of laryngeal and supralaryngeal structures. They also note an alternative interpretation, however. Delayed lip closure is accompanied by a delay in build-up of oral air pressure. Oral air pressure may contribute to voicing offset by reducing the transglottal airflow. Hence the delayed voicing offset may be a mechanical consequence of the delayed lip closure.

3. The laryngel reflex system whose receptors are sensitive to air pressure, according to Wyke (1967), has a respiratory as well as a laryngeal component. An increase in air pressure, as noted, leads to an increase in the tone of vocal fold adductors. In addition, Wyke cites some studies showing that these same receptors control unspecified alternations in the activity of respiratory muscles.

4. Finally, Gould (1971) reports that changes in posture (e.g., standing versus sitting curled in a chair) lead to reflexive alterations in the respiratory and laryngeal musculature.

## SUMMARY

The data described in the four preceding sections are highly compatible with the mode of organization of the nervous system and musculature proposed by Greene and by Turvey. Although the data do not provide a clear picture of how all of the coordinative structures fit together to generate a coherent, homogeneous act of speech production, that is not surprising given that none of them was gathered with a view to supporting this theory of action. Indeed the quantity of the supportive data is remarkable in view of this.

### Coarticulation in an Intrinsic Timing View of Speech Production

The account of coarticulation to be developed here meets its obligations (as listed earlier) in these ways:

1. Phonological segments are defined by the coordinative structures that are invoked in their realization. Since coordinative structures engender four-dimensional acts, phonological segments are considered essentially or canonically four dimensional. Those phonological segments produced by the same set of coordinative structures (e.g., the class of vowels) are distinguished by the parameter values that the muscle systems are assigned. (The parameter values are equivalent, or nearly so, to distinctive features.)

2. Consonants and vowels are distinguished by the coordinative structures that effect their realization. An argument can be made that all vowels are the product of a single set of coordinative structures invoked just once at the onset of an utterance. These muscle systems effect a characteristic kind of gesture (a relatively slow change in the global shape of the vocal tract) that distinguishes vowels as a class of segment. Consonants are produced by (a variety of) coordinative structures different from those that invariantly underlie vowel production. The vowel-producing system may be cyclically invoked, thereby yielding quasi-stress timing in languages such as English.

3. The plan for an utterance is treated as identical to the coordinative structures themselves--that is, with the patterning of physiological biases that organize the musculature. This decision does not preclude explaining anticipatory effects in speech production (that is, speech errors, anticipatory shortening, anticipatory coarticulation). It does not because, as noted, the coordinative structures are nested; the superordinate relationships are established both with respect to what the talker is doing now and with respect to what he will be doing.

4. Coarticulation is characterized as the coproduction of four-dimensional "canonical" forms.

### The Organization of a Talker During Speech

That vowels and consonants are coproduced has been suggested by Kozhevnikov and Chistovich (1965), by Ohman (1966, 1967) and by Perkell (1969). For Kozhevnikov and Chistovich, the initial consonant and the vowel of a $C_0V$ syllable are initiated simultaneously by a talker. Given the absence in running speech of articulatory gaps or pauses, this view implies that vowels are continuously produced. (More accurately, from the perspective of Kozhevnikov and Chistovich, it implies that a second vowel is initiated as soon as a preceding vowel is terminated.) The production of a consonant or a consonant cluster, then, is imposed on a background of continuous vowel production.

This view is stated more explicitly by Ohman (1966, 1967) based on acoustic data showing that the formant transitions into (and out of) a medial stop consonant in a VCV vary with the identity of the following (and preceding) vowel. Again, if this is the case in a VCV, by implication vowel production may be continuous throughout the course of running speech.

Apparent coproduction of vowels and consonants may, but need not, imply that the two kinds of speech gestures are produced by different articulatory systems. A way in which coproduction might be approximated by a single articulatory system is by means of feature spreading. However, I have already argued that this view is implausible. A more plausible way, as suggested both by Ohman and Perkell, is for the articulatory systems responsible for vowel and consonant production to be distinct.

Ohman notes that the global shape of the vocal tract during a stop closure is irrelevant to the phonemic identity of the consonant. It can vary without altering the identity of the stop for a listener. Thus if it is mechanically feasible, a "distorted" vowel gesture (Ohman, 1966) can be executed by the tongue-body during the closures for bilabial consonants, for alveolar consonants, and even, he suggests, for consonants involving the body of the tongue. His acoustic evidence bears this out as does the more recent articulatory data of Butcher and Weiher (1976) and of Barry and Kuenzel (1975).

What makes coproduction mechanically feasible to Ohman and Perkell is that the production of vowels and consonants involves essentially different (but overlapping) sets of muscles. It is perhaps worth quoting Perkell (1969) at some length on this point:

> The behavior of the vocal tract differs in several respects for the production of vowels and consonants. The division of the observed parameters of vocal-tract behavior into classes based on the articulatory and acoustic distinctions between vowels and consonants suggest criteria that can serve as the bases for a physiological model. Many parts of the vocal tract play a role in the production of both vowels and consonants, but in general, the same organs seem to behave differently under the influence of two different classes. Consonant articulations by the tongue and lips are generally observed to be faster and more geometrically complex, and they require more precision in timing than vowel articulations... To some extent there also seems to be an anatomical division. For example, the tongue tip is more active in consonant articulation, whereas the body of the tongue is active in articulating both consonants and vowels.
>
> The general differences in velocity, complexity, precision of movement, and in anatomy suggest that different types of muscles are

> generally responsible for consonant and vowel production. It is probable that articulation of vowels is accomplished principally by the larger, slower extrinsic tongue musculature which controls tongue position. On the other hand, consonant articulation requires the addition of the precise, more complex, and faster function of the smaller intrinsic tongue musculature. (p. 61)

Taken together, these observations, and extrapolations from them indicate that true coproduction occurs in speech, and that the capacity for coproduction derives from an adaptive property of speech that the two classes of articulatory gestures, consonants and vowels, are products of different (coordinated) neuromuscular systems.

This property is adaptive for the following reasons. Any two vowels are more similar in their acoustic form than any vowel is to a consonant. Their acoustic form, of course, is consequent to their manner of production. Vowels are produced as (relatively) slow alterations in the global shape of the vocal tract effected in large part by repositioning the tongue-body in the mouth. This characterization is common to vowels, but is not common to consonants and vowels. If vowels and consonants are produced by different neuromuscular systems, then vowels may be continuously produced as suggested above; those articulatory properties that are common to all vowels are invariant over the course of an utterance. Conceivably they are evoked anew at the initiation of each vowel. However a more parsimonious supposition is that they are evoked just once for the whole course of an utterance.

In other words, the supposition is that the set of articulatory properties that is common to the vowels, and that defines them as a natural class of gestures, may be the product of a superordinate coordinative structure whose time-scale is slow relative to that of the ongoing speech gestures. If vowel production is continuous, then this hypothetical coordinative structure can be evoked just once for the course of an utterance. If instead, vowel production alternates with consonant production, then the talker cannot exploit the properties of vowels that define their equivalence. Rather, he has to reevoke the invariant (as well as the variant) properties of the class of vowels at the initiation of each new vowel. Below I will try to establish what the invariant properties of the class of vowels are, and then to characterize the coordinative structure that evidences them.

By hypothesis, five muscle systems are coordinated to produce a vowel: the first is the respiratory system characterized earlier; the second is the intrinsic musculature of the tongue, which on a first approximation is set invariantly across the vowels (Perkell, 1969); two systems are responsible for adjusting the length of the vocal tract (see Perkell, 1969, pp. 40-41), and finally a system is responsible for moving the tongue within the oral cavity. These systems are characterized in Fowler (1977).

The last vowel system may be described (at least metaphorically) as if it were a vibratory system with its resting length as a tunable parameter. Even a simple vibratory system, for example a linear spring described by the equation $-F = k(l - l_o)$, has some properties not unlike those of the articulatory system during vowel production.

To describe the production of vowels, we need a system for positioning the tongue that captures those properties of the class of vowels that are equivalent across its membership. A spring-like equation with the parameter $l_o$ unspecified would satisfy that criterion because it represents a system that is established invariantly for every vowel. $l_o$ is a parameter whose particular value distinguishes one vowel from another.

In addition, we are looking for a description of a particular vowel that enables us to describe as equivalent the different gestures that instantiate it in different contexts. (Recall that for /ε/, the tongue is lowered following /i/, but raised following /a/). Possibly $l_o$ will work to do that if we describe it at an abstract level as the zero-state of the extrinsic tongue system. l is the actual state of the tongue system (the actual position of the tongue).

Consider what happens to a spring system when F and k are unchanged but $l_o$ is reduced in magnitude. To counteract the same F, the system decreases l by the same amount as $l_o$ was decreased. Thus l alters in the direction of the new $l_o$. (For example, let $l_o = 10$ in arbitrary units, $F = 50$, $k = 25$. Transforming the equation above, $l = 10 - (F/k) = 10 - 2 = 8$. If now $l_o$ is reset to 5, $l = 5 - 2 = 3$.) Again suppose that $l_o$ corresponds to the zero-state of the extrinsic tongue system--to the position that the tongue would adopt if $-F = 0$, and l is the actual position of the tongue. If a talker is able to alter $l_o$ volitionally, as the subjects of Fel'dman could (see above), then he can change the zero-state of the extrinsic musculature. Suppose that to a particular vowel corresponds a particular value of $l_o$. A new vowel is initiated by changing the value of $l_o$. In consequence of this change in the value of $l_o$, the invariant shape of the tongue in the mouth alters its location. In the case of the vowel /ε/ its $l_o$ is less than that for the vowel /i/ and greater than for /a/. When $l_{oe}$ is substituted for $l_{oa}$, the tongue moves upwards; when $l_{oe}$ is substituted for $l_{oi}$; the tongue moves downwards. Although the gesture is different for /ε/ in different contexts, the parameter $l_{oe}$ is the same. And therefore it is incorrect to equate the view of the extrinsic muscle system of the tongue as a spring system with a feature-bundles or targets view. The reason why this is so can be stated in two ways. First, $l_o$ is just a parameter of a system. /ε/ is the functioning of the system when the spring function is assigned the parametric value $l_{oe}$, but /ε/ is not identical with the parameter itself. Equivalently, it is as incorrect to equate $l_{oe}$ with /ε/ as it is to equate some curve with its asymptote. $l_{oe}$ is just the limiting shape of the vocal tract (as controlled by the position of the tongue) towards which /ε/ invariantly aims.

Now consider how this proposal handles some coarticulatory phenomena.

Coarticulatory effects are due to the coproduction of consonant and vowels and of stressed and unstressed vowels. Consider first the coarticulatory effects of vowels on consonants. Lip rounding precedes the measured acoustic onset of a rounded vowel and therefore coarticulates with consonants that precede a vowel. This occurs, we suppose, not because the feature [+rounding] has attached itself in the plan to the preceding consonants, but rather because the vowel /u/ is coproduced with them. Vocal tract lengthening via lip rounding is, then, an observable correlate of co-production. Another correlate, reported by MacNeilage and DeClerk (1969), is that the tongue-body configuration for the vowel may be attained during the closure for a consonant (see also Barry & Kuenzel, 1975; Butcher & Weiher, 1976).

Similarly, velar opening precedes the gesture of the primary articulator for a nasal consonant. This occurrence probably parallels that of lip rounding for the vowels. That is, it may participate in a coordinative structure, the component gestures of which are patterned over time rather than being synchronized. Velar opening occurs during a vowel then, because vowels and consonants are coproduced. Parallel to the finding of Bell-Berti and Harris described earlier, one might expect velar opening to precede the major articulatory gestures for an /m/ or /n/ by a relatively fixed duration, and not to be synchronized consistently with the gestures toward a vowel target.

Coarticulatory vowel-to-vowel effects may again be explained as owing to coproduction. The left-to-right and right-to-left transconsonantal effects observed by Fowler (1977) of stressed vowels on $F_2$ of an intervening unstressed vowel is explained most naturally as coproduction. The production of an unstressed vowel is superimposed on a trajectory of the shape of a vocal tract from one stressed vowel to another (cf. Martin, 1972).

## Concluding Remarks

This section has provided a sketchy and speculative view of intrinsic timing in speech production. However, I believe that it can account for coarticulatory and other timing effects more plausibly and adequately than the views developed within the extrinsic timing framework. Its major advantage is in incorporating the dimension of time into the specification of a phonological segment with the consequence that the ideal or canonical form is considered to be executed unaltered in an utterance (cf. Fowler et al., in press).

## REFERENCE NOTES

1. Abramson, A. Laryngeal timing in consonant distinctions. Paper presented at the Eighth International Congress of Phonetic Sciences, Leeds, England, August 17-23, 1975.
2. Harris, K. Vowel duration change and its underlying physiological mechanisms. Paper presented at the 50th session of the American Speech and Hearing Convention, 1975.
3. Rapp, K. A study of syllable timing. Papers from the Institute of Linguistics, University of Stockholm, November 14-19, 1971.

## REFERENCES

Abercrombie, D. Elements of general phonetics. Chicago: Aldine, 1967.

Allen, G. The location of rhythmic stress-beats in English: An experimental study I. Language and Speech, 1972, 15, 72-100.

Barry, W., & Kuenzel, H. Co-articulatory airflow characteristic of intervocalic voiceless plosives. Journal of Phonetics, 1975, 3, 163-282.

Bell-Berti, F. The velopharyngeal mechanism: An electromyographic study. Supplement to Haskins Laboratories Status Report on Speech Research, 1973.

Borden, G. J., & Gay, T. Durations of articulator movements for /s/-stop clusters. Haskins Laboratories Status Report on Speech Research, 1975, SR-44, 147-161.

Bosma, J. F. A correlated study of the anatomy and motor activity of the upper pharynx by cadaver dissection and by cinematic study of patients after maxillofacial surgery. Annals of Otology, Rhinology and Laryngology, 1953, 62, 51-72.

Butcher, A., & Weiher, E. An electropalatographic investigation of coarticulation in VCV sequences. Journal of Phonetics, 1976, 4, 59-74.

Daniloff, R. G., & Hammarberg, R. E. On defining coarticulation. Journal of Phonetics, 1973, 1, 239-248.

Doty, R. W. Neural organization of deglutition. In C. F. Code (Ed.), Handbook of physiology (Vol. IV, Sec. 6). Washington: American Physiological Society, 1968.

Draper, M., Ladefoged, P., & Whitteridge, D. Respiratory muscles in speech. Journal of Speech and Hearing Research, 1959, 2, 6-27.

Easton, T. On the normal use of reflexes. American Scientist, 1972, 60, 591-599.

Fawcus, B. Oropharyngeal function in relation to speech. Developmental Medicine and Child Neurology, 1969, 11, 556-560.

Fel'dman, A. G. Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II. Controllable parameters of the muscles. Biophysics, 1966, 11, 565-578.

Folkins, J., & Abbs, J. H. Lip and jaw motor control during speech: Responses to resistive loading of the jaw. Journal of Speech and Hearing Research, 1975, 18, 207-220.

Fowler, C. A. Timing control in speech production. Bloomington, Indiana: Indiana University Linguistics Club, 1977.

Fowler, C. A., Rubin, P., Remez, R., & Turvey, M. T. Implications for speech production of the general theory of action. In B. Butterworth (Ed.), Speech production. New York: Academic Press, in press.

Fowler, C., & Turvey, M. T. Observational perspective and descriptive level in perceiving and acting. In W. Weimer & D. Palermo (Eds.), Cognition and the symbolic processes II. Hillsdale, N.J.: Lawrence Erlbaum Associates, in press.

Gay, T., & Hirose, H. Effect of speaking rate on labial consonant production: A combined electromyographic/high speech motion picture study. Phonetica, 1973, 27, 44-56.

Gay, T., & Ushijima, T. Effect of speaking rate on stop consonant-vowel articulation. Haskins Laboratories Status Report on Speech Research, 1974, SR-39/40, 213-217.

Gay, T., Ushijima, T., Hirose, H., & Cooper, F. S. Effect of speaking rate on labial consonant-vowel articulation. Journal of Phonetics, 1974, 2, 47-63.

Gibson, J. J. The theory of affordances. In R. Shaw & J. Bransford (Eds.), Perceiving, acting and knowing: Towards an ecological psychology. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Gould, W. J. Effect of respiratory and postural mechanism upon action of the vocal cords. Folia Phoniatrica, 1971, 23, 211-224.

Greene, P. H. Problems of organization of motor systems. In R. Rosen & F. Snell (Eds.), Progress in theoretical biology (Vol. 2). New York: Academic Press, 1972.

Grillner, S. Locomotion in vertebrates. Physiological Reviews, 1975, 55, 247-304.

Hammarberg, R. The metaphysics of coarticulation. Journal of Phonetics, 1976, *4*, 353-363.

Kaplan, H. M. *Anatomy and physiology of speech* (2nd ed.). New York: McGraw-Hill, 1971.

Kent, R., Carney, P., & Severeid, L. Velar movement and timing: Evaluation of a model for binary control. *Journal of Speech and Hearing Research*, 1974, *17*, 470-488.

Kent, R. D., & Minifie, F. D. Coarticulation in recent speech production models. *Journal of Phonetics*, 1977, *5*, 115-117.

Kent, R. D., & Moll, K. L. Articulatory timing in selected consonant sequences. *Brain and Language*, 1975, *2*, 304-323.

Kent, R., & Netsell, R. Effects of stress contrast on certain articulatory parameters. *Phonetica*, 1971, *24*, 23-44.

Kozhevnikov, V. A., & Chistovich, L. A. Speech: Articulation and perception. Moscow-Leningrad, 1965. (English translation: J.P.R.S., Washington, D.C., No. JPRS 30543.)

Lashley, K. The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior*. New York: Wiley, 1951.

Lenneberg, E. *Biological foundations of language*. New York: Wiley, 1967.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, A., Delattre, P., Gerstman, L., & Cooper, F. S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, 1956, *52*, 127-137.

Liberman, A., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz, & H. L. Teuber (Eds.), *Handbook of sensory physiology* (Vol. VIII), "Perception." Heidelberg: Springer-Verlag, 1978.

Lieberman, P. *Intonation, perception, and language*. Cambridge, Mass.: MIT Press, 1967.

Lindblom, B. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, *35*, 1733-1781.

Lisker, L. On time and timing in speech. In T. Sebeok (Ed.), *Current trends in linguistics* (Vol. XII). The Hague: Mouton, 1972.

MacKay, D. G. Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, 1970, *8*, 323-350.

MacNeilage, P. Motor control of serial ordering of speech. *Psychological Review*, 1970, *77*, 182-196.

MacNeilage, P., & DeClerk, J. L. On the motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America*, 1969, *45*, 1217-1233.

MacNeilage, P., & Ladefoged, P. The production of speech and language. In M. P. Friedman & E. C. Carterette (Eds.), *Handbook of perception* (Vol. 7): *Language and speech*. New York: Academic Press, 1976.

Martin, J. Rhythmic (hierarchical) vs serial structure in speech and other behavior. *Psychological Review*, 1972, *79*, 487-509.

Morton, J., Marcus, S., & Frankish, C. Perceptual centers (P-centers). *Psychological Review*, 1976, *83*, 405-408.

Neisser, U. *Cognition and reality*. San Francisco: Appleton-Century-Crofts, 1976.

Ohala, J., Hiki, S., Hubler, S., & Harshman, R. Photoelectric methods of transducing lip and jaw movements in speech. *UCLA Working Papers in Phonetics*, 1968, *10*, 135-144.

Ohman, S. Coarticulation in VCV utterance: Spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, *39*, 151-168.

Ohman, S. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 1967, *41*, 310-320.

Perkell, J. *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, Mass.: MIT Press, 1969.

Polanyi, M. *Personal knowledge*. Chicago: University of Chicago Press, 1958.

Sessle, B. J., & Hannam, A. G. (Eds.). *Mastication and swallowing: Biological and clinical correlates*. Toronto: University of Toronto Press, 1975.

Shohara, H. The genesis of the articulatory movements of speech. *Quarterly Journal of Speech*, 1936, *21*, 343-348.

Sternberg, S., Monsell, S., Knoll, R., & Wright, C. E. The latency and duration of rapid movement sequences: Comparison of speech and typewriting. In G. Stelmach (Ed.), *Information processing in motor control and learning*. New York: Academic Press, 1978.

Studdert-Kennedy, M. Universals in phonetic structure and their role in linguistic communication. In T. Bullock (Ed.), *Recognition of complex acoustic signals*. Berlin: Dahlem, 1977.

Sussman, H. What the tongue tells the brain. *Psychological Bulletin*, 1972, *77*, 262-272.

Turvey, M. T. Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Towards an ecological psychology*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Wickelgren, W. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 1969, *76*, 1-15.

Wyke, B. Recent advances in the neurology of phonation: Phonatory reflex mechanisms in the larynx. *British Journal of Communication Disorders*, 1967, *2*, 2-14.

Wyke, B. Laryngeal myotactic reflexes and phonation. *Folia Phoniatrica*, 1974, *26*, 249-264.

## FOOTNOTES

[1] In an extrinsic timing theory, the articulatory plan represents the serial ordering of features, segments or syllables--that is, it represents their ordinal relationships along the time axis--but time is not taken to inhere in, or to be essential to, the specification of these production units.

[2] It is perhaps worth pointing out that Hammarberg makes two separate claims here: first that there are no invariant correlates of a segment in an acoustic or articulatory record of an utterance, and second that it takes a human mind to perceive a segment. In the quoted passage, Hammarberg evidently makes both claims, but elsewhere (see his discussion of inherent and derived properties of a segment), he seems only to make the second. It is the case that the second claim could be true without the first more radical claim also being true. For example (see Fowler & Turvey, in press, for an elaboration of this point), it probably takes a human mind to recognize something as being an example of "footwear." The essential properties of footwear--e.g., that it be foot-sized and shaped, that it have a means of attachment to the foot and so on--clearly are given in the optical signal to an eye. However, that collection of properties is only a significant collection for an animal that wears things on its feet. Other animals will not recognize that aggregate of

properties as a significant collection and hence will not detect that a shoe, for example, is footwear.

Likewise, if it is true that it takes a human perceptual system to perceive a segment (or a human knowledge system to know one), it need not also be true that the percept is based on properties that are absent in the acoustic signal. I will suggest later that the properties cannot reasonably be supposed to be absent.

[3] The equation for a linear spring is $-F = k(l-l_o)$ where $l_o$ is the spring resting length (that is, the length of the spring in the absence of any force exerted on it). $l$ is the current length of the spring; $k$ is the stiffness parameter; and $F$ is the force developed by the spring. Turvey (1977) suggests that all coordinative structures may be vibratory systems, though not necessarily linear.

# ORTHOGRAPHY AND THE BEGINNING READER*

Isabelle Y. Liberman,+ Alvin M. Liberman,++ Ignatius G. Mattingly+ and Donald Shankweiler+

## INTRODUCTION

We are happy to be members of this Conference, the more so because our research into the reading of English has led us to appreciate how enlightening it might be to examine the reading of other languages and other orthographies. Our aim, then, is to view our research in that light which best reveals just what it is that studies by colleagues in other lands might teach us.

Most of our research has been concerned with the processes and problems that occur in the beginning reader. It divides quite naturally into two parts. One deals with the importance to the reader of having some degree of sophistication about the linguistic structures that the orthography represents, and with the difficulty that attends the development of such sophistication in many beginners. While the importance of that sophistication is fixed, the difficulty of achieving it ought to vary greatly with the nature of the orthography and also, though perhaps less obviously, with the relation of the orthography to certain characteristics of the language. The other part of our research has to do with the importance to the reader of recovering a phonologic representation of that which he reads, especially for the purpose of meeting the short-term memory requirements that language imposes on those who would store the words long enough to understand the sentence. Since all languages impose that requirement--the meaning of a sentence is always distributed among the several words it comprises--we should expect that the results we have obtained with English would apply universally, but it remains to be determined whether, in fact, they do.

## LINGUISTIC SOPHISTICATION: PROBLEMS OF THE BEGINNING READER THAT MAY VARY ACROSS LANGUAGES AND ORTHOGRAPHIES

The point of departure for our earliest research on the tribulations of the beginning reader was the assumption that we were, after all, asking him to

*Paper presented at Cross-Language Conference on Orthography, Reading, and Dyslexia, sponsored by NICHD and Fogarty International Center of NIH, Bethesda, Maryland, September 18, 1978.

+Also University of Connecticut.

++Also University of Connecticut and Yale University.

Acknowledgment: The preparation of this paper, as well as the research on which it is based, was supported by grant HD01994 from the National Institute of Child Health and Human Development.

do something quite unnatural. That assumption appeared to us obvious, if only because reading and writing seem rather far removed from their biological roots in the universals of language. We know that reading and writing appear late in the history of humankind, just as they do in the development of the individual; and also that there is considerable variation among orthographies in the nature and size of the linguistic units (phonemes, morphophonemes, syllables, moras, morphemes) they represent. We therefore supposed that the (less natural) processes of reading and writing would need to be more deliberate than the (more natural) processes of listening and speaking. In particular, we put our attention on the possibility that, in contrast to the listener and speaker, the reader and writer must be a kind of linguist. The largely tacit command of language that serves the nonlinguist, when, in speaking and listening, utterances roll trippingly off his tongue or pass readily into his comprehension, is not sufficient for the reader and writer; like the linguist, he requires a greater degree of sophistication about linguistic structures, including, in particular, those that are represented by the orthography he reads or writes (Gleitman & Rozin, 1977; I. Y. Liberman, 1971, 1973; I. Y. Liberman, Shankweiler, A. M. Liberman, Fowler, & Fischer, 1977; Mattingly, 1972).

The sophistication that is required has two aspects, corresponding approximately to two aspects of the way an orthography represents speech. The one, which we will call "phonological maturity," has to do with the often abstract but nonetheless regular nature of the link between the orthography and the phonetic (or phonemic) structures it conveys. In English, for example, the spellings of words such as _telegraph_, _telegraphy_, and _telegraphic_ are irregular except as the reader comprehends the (morpho-)phonological rules that rationalize them. Phonological maturity is, as in the case just cited, of some importance to the beginning reader, though it is not, in our view, crucial. More important by far is an explicit understanding by the reader of the relation in segmentation between the orthography and speech. It is patent that an alphabet, for example, can be used properly only if the reader (and especially the beginner) is quite aware that speech is divisible into those phonological segments that the letters represent. This aspect of sophistication about language we will refer to as "linguistic awareness" (I. Y. Liberman, 1971, 1973; Mattingly, 1972).

In this section, we will offer our views about the roles of these two aspects of linguistic sophistication, summarize our research on the development of linguistic awareness and on its relation to success in reading, and, finally, speculate about the interaction between phonological maturity and linguistic awareness, on the one hand, and, on the other, the orthography and the language.

### The Role of Phonological Maturity in Learning to Read

A reader is able to recognize a written word because he can equate it with some representation of that word stored in long-term memory. We believe that this stored representation is linguistic, and that an orthography appeals to the reader's appreciation of the grammatical structure of utterances. Specifically, we begin with Chomsky's (1970) argument that the orthographic transcription of a word corresponds approximately to the way generative phonologists assume the word is represented in the ideal speaker-hearer's mental lexicon. This representation is often morphophonological: the word is conveyed as a sequence of systematic phonemes divided into its constituent morphemes. For example, the words *heal*, *health*, *healthful*, have the morphophonological representations[1] /hēl/, /hēl+θ/, /hēl+θ+ful/, respectively.

The morphophonological representation of a word is quite distinct from its phonetic representation--that is, from what the speaker-hearer thinks he pronounces and perceives. In the phonetic representation, *heal* and *health* are realized, approximately, as [hīyl] and [helθ]. Notice that in the phonetic representation, the underlying morphophonological forms are to a considerable extent disguised, and explicit morpheme boundaries are absent. Moreover, the same morpheme has various phonetic representations depending upon the phonological context (Chomsky & Halle, 1968).

Clearly, the transcriptions of *heal* and *health* in English orthography approximate the morphophonological representations rather than the phonetic. The orthographic forms differ from the morphophonological representations only in the omission of morpheme boundaries and in the conventional substitutions of *ea* for /e/ and *th* for /θ/.

Chomsky's argument about the morphophonological nature of orthographies applies, of course, to logographic and syllabary scripts as well as to alphabetic scripts. Since English is written alphabetically, we use a distinct symbol for each of the distinct systematic phonemes: /h/, /ē/, /l/, and so on. If English were written logographically, we would use a distinct symbol for each of the morphemes /hēl/, /θ/, /ful/; if it were written in a syllabary, we would use a distinct symbol for each of the syllables /hēl/, /hēl+θ/, /ful/. But in all cases, we would be transcribing the morphophonological representations.

An orthography makes the assumption that readers know, tacitly, the phonology of the language, so the representation of words in their personal lexicon matches the transcriptions of the orthography. In our example, English speakers have the morphophonological representations /hēl/, and /hēl+θ/ in their lexicons--and not [hīyl] and [helθ]. In the course of acquiring English, they have mastered the morphophonological rules, and have inferred that [hīyl] and [helθ] can both be derived from /hēl/, /θ/ being a separate morpheme.

Thus, to the extent that English is written morphophonologically, then to that extent it assumes an ideal reader who commands the grammatical rules in terms of which the spelling makes sense. That is, it assumes a reader who has achieved what we have chosen to call phonological maturity. To a reader who lacks that maturity, the linguistic regularities that justify the orthography are simply opaque, and the spellings can only appear exceptional.

Research by various psycholinguists indicates that young children are, in fact, quite immature phonologically, hence not well equipped to take maximum advantage of the morphophonological aspects of English orthography. Rather, they appear, as speaker-hearers, to learn enough to permit pragmatic communication and only later, if at all, to approach the phonological competence of the ideal speaker-hearer (Berko, 1958; Moskowitz, 1973; Read, 1975). Moreover, there is evidence that, given free rein to spell as they will, such children tend to be rather better as phoneticians than they are as phonologists (Read, 1975; Zifcak, 1977). If so, and if, indeed, a morphophonological orthography is, as some claim, the best one for adults, then English puts the child at odds with the adult.

It is fortunate, therefore, that, while phonological maturity may be of some importance in reading, it is, in no sense, critical. That is, it appears that children who are more at home with a phonetic structure than with a morphophonological one can nevertheless learn to read. At all events, their problem could certainly be minimized by controlling the vocabulary used in early reading instruction. Moreover, informal observation and some experimental evidence suggest that the experience of reading itself serves to stimulate phonological development. Thus, Moskowitz (1973) has shown that a by-product of learning to read is that the child is led to acquire the Vowel Shift rule.

Children who profit from the linguistic stimulation of reading, internalizing the phonological rules they induce from orthographic transcription, and accordingly revising the representations of words in their lexicon to make them more nearly morphophonological, are the sort who continue the process of language acquisition far beyond the pragmatic level. Obviously, they cannot do this except as they read "analytically"--that is, with attention to the relation between the internal structure of the printed word and the phonology of the spoken word. But, given that strategy, they are likely to become more competent users of their language and also, we should suppose, superior readers.

## The Role of Linguistic Awareness in Learning to Read

So much, then, for the difference between a morphophonological representation and a phonetic one, and for the phonological maturity that enables a sophisticated reader to bridge the gap. We turn now from that gap to one that yawns equally wide and presents a much greater hazard for the beginning reader. For if orthographies are morphophonological rather than phonetic, they are *a fortiori* not acoustic or auditory. Though closer to the speech signal, the phonetic representation is far from isomorphic with it. To bridge the gap between the phonetic level and sound, the reader needs what we have called linguistic awareness. To see just what that is, and why it might be hard to achieve, we should consider first one of the peculiar complications that characterizes the relation between phonetic structures and their acoustic vehicles.

Given the way speakers articulate and coarticulate, the segments of the phonetic structure do not correspond in any direct way to the segments of the sound. Thus, a word like *dog* that has three phonological (and orthographic) segments has only one isolable segment of sound (A. M. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). The information for the three phonologic segments is there, but so thoroughly overlapped (encoded) in the sound that there is no way to divide the sound into segments so that each acoustic segment carries information about only one phonetic segment. Nor is the opposite possible. That is, we cannot begin with prerecorded sounds for each of the three segments that we write as *d*, *o*, and *g* and in any way put them together to form the word /dɔg/. An obvious consequence is that many of the segments--in particular, many consonants--cannot be produced in isolation, as syllables and words can; hence these segments might be expected to have little salience and to escape the conscious awareness of the ordinary user of the language.

This characteristic of speech offers no obstacle to the listener, because all speaker-hearers of a language, even very young children, are presumably provided with a neurophysiology that functions quite automatically--that is, below the level of awareness--to extract phonetic structure from the continuous acoustic signal in which it is so peculiarly encoded (A. M. Liberman et al., 1967). To understand a spoken utterance, therefore, the child need not be explicitly aware of its phonetic structure any more than he need be aware of its syntax. But that explicit awareness of phonetic structure is precisely what is required if the beginning reader is to take full advantage of an alphabetic system of writing.

Returning to our example of the word _dog_, consider the child who, knowing the word, sees it in its printed form for the first time. In mapping the three letters onto the word he already knows, it will avail him little to be able to recognize the three letters, and to "sound them out." He must also be consciously aware that the word he knows has three phonetic segments. Without that awareness, and given the impossibility of producing the phonetic segments in isolation, the best the child can do is to say something like [də] [ɔ] [gə], thus producing a nonsense trisyllable that bears no certain relationship to the word /dɔg/.

Indeed, neither the child nor any other reader can recover speech from print on a letter-by-letter basis. Rather, he must group the letters so as to have put together just those strings of phonetic segments that are, in the normal processes of speech production, collapsed into a single coding unit. (A syllable is sometimes thought to be such a unit.) But there is no simple rule by which a reader can do this. The properly "speakable" unit may comprise almost any number of letters from one to nine or, at the level of prosody, even more. We suspect that acquiring the ability to do this--that is, knowing how to combine the letters into units appropriate for speech--is an aspect of reading skill that, as much as any other, separates the fluent reader from the beginner who has only just succeeded in discovering what an alphabetic orthography is all about (I. Y. Liberman et al., 1977).

Considerations of the kind we have just reviewed led us to suppose that linguistic awareness--awareness of the phoneme in the case of an alphabet--might be difficult for young children, but also important if they are to become readers. In the sections that follow, we will summarize two strands of our research that bear on these suppositions. One deals with the development of linguistic awareness in young children, the other with its relation to reading.

## _Development of Linguistic Awareness_: _Some Experiments_

Given the way most consonant phones are encoded in the sound, it is, as we have pointed out, not possible to produce them in isolation.[2] But syllables can be so produced. (Vowels can, of course, be treated as if they were syllables.) We should suppose, then, that it might be easier for the child to become aware of syllables than of phonemes.[3] Indeed, we might even suppose, more generally, that this difference accounts for a fact about the history of writing systems--to wit, that syllabaries appear early and as a result of several quite independent developments, in contrast to an alphabet, which appears later and only once. Looked at this way, the alphabet can be seen as a triumph of applied linguistics, a cognitive achievement by the race. Is it so for the child, too? To find out, we and our colleagues have carried out experimental studies designed to compare the development in the child of awareness about syllables and phonemes.

The object of the first experiment (I. Y. Liberman, Shankweiler, Fischer, & Carter, 1974) was to compare the ability of children in nursery school, kindergarten, and first grade (four-, five-, and six-year-olds) to count the phonemes in spoken utterances with the ability of matched groups of children to count syllables. The procedure was in the form of a game that required the child to repeat a word spoken by the experimenter and then to indicate, by tapping a wooden dowel on the table, the number of segments in the word. In order to teach the child what was expected of him, the test list was preceded by a series of demonstration trials. The test proper consisted of randomly assorted items of one, two, or three segments, presented without prior demonstration and corrected, as needed, immediately after the child's response. Testing continued until the child reached a criterion of tapping six consecutive items correctly, or until the end of the list.

It was immediately apparent from this experiment that syllables were more readily counted than phonemes. The number of children who reached criterion was markedly greater in the syllable group, whatever the grade level. None of the nursery school children and only 17 percent of the kindergarteners could count phonemes, while 46 percent of the nursery school children and 48 percent of the kindergarteners could count syllables. The first graders performed much better on both tasks, but only 70 percent could count phonemes while 90 percent were successful with syllables. Similar results have been found with different subject populations in two other investigations by our research group (Treiman, 1976; Zifcak, 1977). We will have more to say about them later. At this point, suffice it to say that in all these studies it was found that explicit analysis of spoken utterances into phonemes is significantly more difficult for the young child than analysis into syllables, and it develops later.

Although awareness of syllables was found to be greater for young children than awareness of phonemes, it was also true that both increased over age, with the steepest increase occurring in the six-year-olds. As it happens, that is the age at which the children in our schools begin to receive instruction in reading and writing. The question immediately arises whether these measured increases represent maturational changes or the effects of experience in learning to read. We will defer discussion of this point to the next section. For now, we should note that, whatever the effects of instruction, our findings strongly suggest that a higher level of linguistic awareness is necessary to achieve the ability to analyze words into phonemes than into syllables.

## Linguistic Awareness and Success in Learning to Read

The argument that linguistic awareness is an important condition for reading has been based thus far on an appeal to sweet reason: it has seemed to us patent that a reader must have explicit knowledge of (at least) the linguistic units that the orthography represents, else he cannot read properly. We should now consider such other bases of support as the argument may have. There are two, both empirical in nature. One has to do with the actual correlation between awareness of segments and success in reading, and also with the possibility that this correlation reflects a causal connection of some kind. The other deals with tests of the correspondence between the errors that beginning readers make, and those we should expect them to make, given the assumption that they are caused in significant measure by the lack of linguistic awareness as revealed by the studies reported in the previous section.

Correlational studies. Recall, now, the gross correlation, reported in the previous section, between the spurt in awareness of phonemic segmentation and the onset of reading instruction. One interpretation of that correlation is, of course, that both are related to age but not to each other. In this connection, we do indeed suspect that age is important for linguistic awareness *and* reading because, being cognitive achievements of sorts, both must require the attainment of some level of intellectual maturity. But, as we have so often implied, we also suspect that the relation between the two is causal, though in a reciprocal way: the awareness, we believe, is important for the acquisition of reading; at the same time, being taught to read helps to develop the awareness.

Let us consider, first, the possibility that linguistic awareness is necessary for reading. Obviously, we should like here to be able to report the results of experiments which show, other things equal, the effects on reading achievement of various kinds of training in awareness of segmentation. Unfortunately, no carefully controlled studies of that kind have been completed, or, at least, none that we know of. Such data as we have are only correlational, but they are, nevertheless, encouraging.

We were motivated to initiate the correlational studies by a rough check of the reading achievement of the group of first graders who had taken part in our experiment on phoneme counting. Testing them at the beginning of their second school year, we found that there had been no failures in phoneme counting among the children who now scored in the top third of the class in reading; in contrast, half of the children who tested in the lowest third of the class in reading achievement had failed in the phoneme counting task the previous year (I. Y. Liberman et al., 1977).

Three subsequent studies by our research group (Helfgott, 1976; Treiman, 1976; Zifcak, 1977) have now substantiated these results. The consistency of positive findings in all these correlational studies, despite widely diverse subject populations, school systems, and measurement devices, gives us confidence that there is, at least, a correlation between awareness of segmentation and success in learning to read.

What, then, of the possibility that instruction in reading is important in the development of linguistic awareness? Here, there is one study that is both relevant and interesting. As yet unpublished, it is by Morais, Cary, Alegria, & Bertelson (1978), who have made a draft version of their paper available to us. These investigators took advantage of a kind of experiment created by particular conditions of life in Portugal. There, they were able to compare awareness of phonemic segmentation in two groups of reasonably matched adults, one illiterate, the other literate. The finding was that the illiterates failed the awareness test and the literate subjects passed, from which the investigators concluded that awareness of phonemic segmentation does not develop independently of instruction in reading. Assuming the generality of that conclusion, we are encouraged to believe that the connection between awareness and reading is not accidental.

Analysis of error patterns. It has seemed reasonable to us that the errors a beginning reader makes might enlighten us about his problems, including those that pertain to linguistic awareness, so we have conducted studies designed to make the appropriate observations (Shankweiler & I. Y. Liberman, 1978; Fowler, I. Y. Liberman, & Shankweiler, 1977; Fowler, Shankweiler, & I. Y. Liberman, in press). We will here review the study (Fowler et al., 1977) that is more directly relevant to our assumption about the relation between linguistic awareness and reading.

In that study, second, third, and fourth graders were asked to read aloud from lists of monosyllabic words in which the position (within the word) of consonant and vowel letters was systematically varied. The children's errors were noted and examined, with particular attention, first, to the effect of position on the likelihood that a particular segment would be misread. A clear pattern emerged. Consonants in final position were consistently misread about twice as often as those in initial position. Though the frequency of all consonant errors dropped markedly from the second through the fourth grade, the two-to-one ratio of errors on final and initial consonants was maintained. Vowels yielded a very different result in that errors were independent of position, and that, too, was found in all three grades.

We can hardly claim that the pattern of errors just described falls inevitably out of our hypothesis about linguistic awareness, but we can see that the pattern and the hypothesis are, nevertheless, nicely consistent. Consider the fact that initial consonant errors are less frequent than final consonant errors, and assume a child who does not explicitly understand the segmentation of the words he speaks. Being able to recognize the letters, and knowing (presumably) that he should go from left to right, he begins with the initial consonant. But, lacking the ability to be sufficiently aware of the segmental structure of the word, and failing, therefore, to appreciate its relation to the structure of its orthographic representation, he cannot properly link the initial consonant to the segment represented by the letter that follows. What he often does then is to produce a word that has the same initial consonant but otherwise bears no particular resemblance to the word he is trying to read. Thus, given the word, dog, he might say [dʌmp]. That procedure will give him a relatively high score on initial consonants, but a low score on succeeding ones.

Consider, now, the opposite findings with vowels, which was that errors were independent of position in the syllable. That, too, makes some sense in terms of our hypothesis. Recall that children find it relatively easy to count spoken syllables, presumably because the syllable (usually) has a vocalic nucleus and a corresponding peak of perceived loudness. Of course, a vowel is the essential part of the vocalic nucleus, and, for that reason, a vowel can be a syllable (as most consonants cannot); hence it can be produced in isolation. We should not be surprised, then, that such difficulty as the child might have with the vowels would not depend on their locations.

There were two other results of the error studies that we will briefly note here, though we are unsure of their relevance (if any) to linguistic awareness and its role in reading. One of these, and the one that appears to be the less relevant, was that the consonant errors tended significantly to take the form of incorrect assignment of one segmental feature; that did not appear to be the case with the vowels. The other result was that the vowel errors were more numerous than the consonant errors, and by a considerable margin. That result lends itself to many possible interpretations, some of them interesting from our point of view and some not. Thus, we must consider that the most egregious irregularities of English spelling seem to be concentrated in the vowels, as in precede and proceed. But some of the regular phonologic alternations lie there too--for example, heal-health--and, as we pointed out in an earlier section of the paper, beginning readers may lack knowledge, either explicit or tacit, of these. Finally, there is the possibility at least that vowels cause more trouble because, when produced (and perceived) in isolation, they are less nearly "categorical" than consonants (A. M. Liberman et al., 1967). To decide among these interpretations will require a great deal more research.

So much, then, for the relation between the pattern of errors in the beginning reader and our hypothesis about the importance, in reading, of a conscious awareness of at least some aspects of linguistic structure. But we cannot close this section without pointing out that the results of the error analysis emphatically support a hypothesis more general than, and basic to, the one about linguistic awareness--namely, that the problems of the beginning reader are primarily cognitive and linguistic, not visual or perceptual. Note the consistency with which the children's errors distinguish consonants and vowels: errors on consonants, but not on vowels, depend on position in the syllable; errors on consonants, but not on vowels, tend to be by segmental feature; and finally, errors on consonants are, by far, the less numerous. It is hard to see how such findings can be accounted for on the assumption that the child is having difficulty in the visual or, more broadly, perceptual sphere. Though we may have less than perfect confidence that our finger has pointed to the exact sources of the difficulty, we can be reasonably sure that, being oriented toward cognition and language, it is, at least, aimed in the right direction.

## The Interaction of Phonological Maturity and Linguistic Awareness with the Nature of the Language and the Orthography

Orthographies vary considerably in the demands they make on the beginning reader. This variation has two essentially independent aspects: first, the depth of the orthography, its relative remoteness from the phonetic representation; and second, the particular linguistic unit--morpheme, syllable or phoneme--that is overtly represented. A deep orthography, like that of English, demands greater phonological development on the reader's part than a shallow orthography, like that of Vietnamese. Logographies (such as the Chinese writing system), syllabaries (such as Old Persian cuneiform), and alphabetic systems (such as that of English) demand successively increasing degrees of linguistic awareness. Neither sort of orthographic variation is to be attributed to historical accident alone: the structure of the language, and perhaps political and social factors, are typically involved. Moreover, advantages for the beginning reader with respect to the phonological maturity or linguistic awareness demanded are often offset by disadvantages of other kinds.

Orthographic depth depends upon two variables: the depth of the morphophonological representation itself and the degree to which the orthography approximates this representation. If the morphophonological representation is quite close to the phonetic representation, the orthography will, of course, be close as well. The reader needs to know little phonology because there is little phonology to be known. This seems to be the case not only with Vietnamese but also with Turkish and many other languages. In the case of Turkish, the orthography is even shallower than the morphophonological representation because the alternations determined by the Vowel Harmony rule (which is about all there is to Turkish phonology) are nevertheless transcribed in the orthography. It can be argued that this is not unreasonable because there are numerous borrowed words which are not subject to Vowel Harmony (A. Kardestuncer, unpublished ms). (By contrast, English orthography transcribes the underlying forms of vowel-shifted words, despite a great many borrowed words that are not subject to Vowel Shift). The orthographies of languages with limited phonologies ought, in general, to be easy for the beginner.

If the morphophonological representation of language is relatively deep, various compromises with the ideal may be observed in its orthography; in particular, phonologically predictable alternations may be explicitly indicated. We have already given some examples from English orthography. In Sanskrit, the alternations between aspirated and unaspirated stops (Grassman's law) are transcribed. In Spanish, infinitives are transcribed without the underlying, phonologically deleted, final /e/ of the morphophonological representation, e.g., /decire/, "to say," is written _decir_ (Harris, 1969). In this respect, as in many others, French orthography, which has _dire_, is closer to the morphophonological representation. The orthography of Spanish, on the other hand, has a surface regularity that accounts in part for its reputation as an "easy" language among American secondary-school students. If a language has an exceptionally deep phonology, it may well be the case that few native speakers actually control very much of it. It is reported that when a morphophonological orthography was devised for Mohawk, native speakers could not learn to use it, and a much shallower orthography had to be substituted (M. Mithune, personal communication).

To make clear that the depth of the orthography is independent of the unit of representation, it may be pointed out that the kana symbols of the hiragana syllabary used for Japanese represent morphophonological syllables, that is, moras. Thus, the kana for a syllable beginning with a voiceless stop is used even when the stop occurs in noninitial position, and so becomes voiced by phonological rule. Moreover, a two-mora sequence, e.g., su ku, will be transcribed with two kanas even though, in colloquial speech, it will often be realized phonetically as [sku]. Thus, the kana, which are usually learned by Japanese children by the time they enter school (Sakamoto, this Conference), require at least a modest degree of phonological maturity. As for linguistic awareness, we should wonder whether moras or phonetic syllables are more readily available.

Languages with deep morphophonological representations appear to put the phonologically immature learner at odds with the more experienced and phonologically more mature reader. An orthography practical for the former may be cumbersome for the latter. But if we are correct in our emphasis on the contribution of reading to phonological maturity, a shallower orthography may reduce the reader's opportunities for learning more about his language.

We turn now to the advantages and disadvantages of transcribing linguistic units other than phonemes. In the case of Chinese writing, the use of a morphemic transcription has a number of advantages. The most obvious, from our point of view, is that it presumably makes minimal demands on linguistic awareness, for, to the extent that morphemes can be produced in isolation, they are salient and readily available to consciousness. In this connection, we should wonder if some difficulties nevertheless arise whenever the phonology makes more abstract the basis for recognizing morphemic identity across words. At all events, the availability of the units is not the only advantage. The various dialects of Chinese can use the same writing system, even though they have developed independently to such an extent that the systematic phonemic representation of a given morpheme will, in general, differ from dialect to dialect. Since the morphemes are, in general, monosyllabic, and since constraints on syllable structure permit only some 1200 phonemically distinct syllables, a syllabary or an alphabetic system would entail substantial homography; this is avoided by the use of a logography. The price, obviously, is that the learner must devote several years to memorizing two or three thousand characters. Having acquired this basic stock, however, he can read a great many more words, since compounding is the basic method of word formation: Chinese content-words are ordinarily bimorphemic (Martin, 1972). In regularly written Japanese text, the kanji logograms are used for roots and the hiragana only for affixes. Thus the Japanese child, like the Chinese child, must devote years to the memorization of characters. The use of kanji, it is said, serves to avoid the homography that would result from a syllabic or phonemic transcription of an almost intolerably homonymic language. The kanji themselves, however, are typically homographic (Martin, 1972).

Syllabary systems are best suited for languages in which the number of possible syllables is small, as in the case of Old Persian, Hittite, and the classical Semitic languages (Gelb, 1963). Semitic had the further advantage that its root morphemes, which were relatively few in number, had the patterns C_C_ or C_C_C_, the intervening vowels carrying only inflectional and derivational information. In the Semitic syllabaries, each symbol stood for any one of the set of CV syllables beginning with a particular consonant. Thus an inventory of only 22 symbols was required, yet a word could be transcribed by only two or three symbols. This resulted in an extremely compact transcription that did not require the reader to be aware of phonemes. But, of course, he had to guess which of the many inflectional and derivational forms of each word was intended and this must have required both control of the complex morphology of Semitic and a keen awareness of it. Evidently this burden was not always endurable, since the practice of using supplementary symbols to disambiguate vowel quality arose early (Gelb, 1963).

From these examples, we might conclude that syllabaries and logographies are realistic possibilities only under rather special linguistic circumstances, and that, even then, the price may be high. For the modern Indo-European languages, which have fairly elaborate syllable structures, large and rather inefficiently exploited inventories of morphemes, and little homonymity, an alphabetic system is preferable, despite the requirement of a relatively greater degree of linguistic awareness.

## PHONOLOGICAL RECODING: A PROBLEM OF THE BEGINNING READER THAT MAY BE MORE OR LESS INDEPENDENT OF LANGUAGE AND ORTHOGRAPHY

One of the advantages of the alphabetic writing system is that, in the ideal case, one can read words he has never before seen. It is obvious, however, that one can do this only insofar as he is able to map the internal structure of the written word onto the segmental structure of the morphophonological representation of the spoken word he holds in his personal lexicon. This requires, as we have said earlier, a degree of linguistic sophistication that many beginning readers do not have and find difficult to attain. If such beginners read at all, they must read holistically. If they do, there are two possibilities. They may be locating the lexical entry by recovering the morphophonological representation as if it were an arbitrary paired-associate of the orthographic transcription, just as the reader of a logographic system must do. Or they may be recovering some sort of semantic representation, attempting to go "directly to meaning." But if the latter is the case, then they stand to lose two advantages that the morphophonological representation affords the readers of all orthographies.

The first advantage relates to lexical lookup, the second to the interpretation of the sentence. We believe it is important that the reader locate the lexical entry for the very word intended by the writer, so that the grammatical and semantic features peculiar to the word are available for subsequent sentence processing. Not everyone appears to concede this; there are some who seem to believe that readers do, or should, read the way aphasics are said to listen, relying heavily on _a priori_ knowledge and common sense, and using the word in the text to narrow down the semantic possibilities a bit, or to suggest some semantically-related word. But if it is granted that the intended word is required, the morphophonological representation provides the most direct means of lexical lookup. Despite minor problems caused by homonymity, a search of the lexicon based on the morphophonological representation is rapid and self-terminating; either the word is there, properly specified, or it is not. This is obviously untrue of a search based on semantic information; how can the "semantic" reader know when he has found the most likely part of the conceptual forest or located the most plausible tree? It was exactly this difficulty, we suppose, that made picture-writing unsatisfactory. Is it also, perhaps, this difficulty that lies behind the tendency of some young readers--presumably those who do not recode phonologically--to land in the right semantic area but on the wrong word, as when, for example, on being shown the word dog, the child reads "cat"?

Note also that, in listening, a normal nonaphasic person locates the lexical entry by what might seem rather a roundabout process: he recovers the phonetic representation by means of the mechanisms of speech perception, and then, either through analysis-by-synthesis, or, more likely, by using various shortcuts, determines what morphophonological representation would generate the phonetic representation consistent with the phonological rules he commands. Then he searches for the lexical entry that corresponds to this morphophonological representation. If Nature seems to find this cumbersome procedure preferable to "going directly to meaning" from the acoustic waveform, and has endowed us with the necessary special-purpose equipment to make the procedure workable in real time, it must be, in part, because of the virtues of the morphophonological representation as a means of locating a lexical entry.

In comparison with this account of the apparently complex processes that go on in understanding speech, the proposal that reading exploits morphophonological representations seems quite straightforward. And at any rate, since speech is prior to reading, the beginning reader has at his disposal a well-established and natural device for lexical lookup. Would it not be disadvantageous for him to set up an entirely new one, and unparsimonious for us to suppose that he must?

The second advantage of the morphophonological representation has to do with its relationship to the nature of the working memory that stores words long enough to permit the sentence they form to be interpreted. We assume, as just indicated, that in the case of speech understanding, morphophonological representations are inferred in working memory from an input representation that is phonetic. It is an important and unsettled question, but one not relevant for our present purpose, whether, in reading, the working memory is essentially morphophonological, or whether a phonetic representation is generated as well even though it would appear to be redundant (Mattingly, 1972). What is relevant is whether, in reading as in speech, a working-memory representation, identical either with the morphophonological representation or with one of its phonetic derivatives, is used--a representation that we shall call, more for convenience than for precision, "phonological."

In speculating about the working memory of a reader, we must consider that some nonphonological representation--visual or semantic, perhaps--might be invoked. Surely, such a strategy is possible--indeed, there is evidence that a visual representation is employed by some congenitally deaf readers--but, as with the matter of lexical lookup, we should suppose that its use is inadvisable.

In any case, there is evidence that, in the case of the normal adult, the nonphonological strategy is not very common. We have in mind several relevant experiments. In some (Conrad, 1964, 1972; Baddeley, 1966, 1968, in press), where information was presented as printed letters, words, or syllables, it was consistently found that the confusions in recall were much greater when the items were phonologically similar than when the similarity was either visual or semantic. This suggests that the readers are storing the information phonologically, though it be disadvantageous to do so. Even when the information is presented in logographic form, strikingly parallel results are obtained. Here, some experiments used Japanese subjects reading the kanji (Erickson, Mattingly, & Turvey, 1973); others had to do with the reading of Chinese (Tzeng, Hung, & Wang, 1977). Finally, the strength of the tendency toward a phonological representation in working memory is underscored by the finding that even when the material presented is not linguistic at all, but pictorial, the information is nevertheless recoded into phonological form (Conrad, 1972). All these results support the idea that use of a phonological representation can be viewed as a generally appropriate strategy for holding linguistic information, however presented, in short-term store.

In view of the memory requirements of the reading task, and evidence for the normal involvement of a phonological representation in the service of that requirement, we were interested in learning whether those beginning readers who are progressing well and those who are doing poorly might be distinguished by the degree to which they rely on a phonological representation when working memory is stressed. We assumed that good beginning readers of an alphabetic orthography, having already related the printed word to the corresponding morphophonological representation, would have the word available for use in working memory in phonological form. Presumably, they would take advantage of that. As for the poor readers, we know that many have difficulty in going the analytic, phonological route and might tend, therefore, to forgo phonological strategies, relying more heavily, perhaps, on representations of a visual or semantic sort.

At all events, we thought it wise to determine whether, in fact, good and poor readers do differ in the degree to which they use a phonological representation in working memory. To that end, we carried out several experiments with children in the second year of the elementary school. In the initial experiments (I. Y. Liberman et al., 1977) we borrowed a procedure devised by Conrad (1972) for adults in which the subject's performance is compared on recall of letters with phonologically confusable (rhyming) and nonconfusable (nonrhyming) names. Our expectation was that the rhyming items would generate confusions and thus penalize recall in subjects who use a phonological representation. Poor readers might then be expected to be less affected by the phonological similarity of the items than good readers, whether or not the groups differed in recall of the nonconfusable items.

The results showed that, though the superior readers were better at recall of the nonconfusable items, their advantage was virtually eliminated when the stimulus items were phonologically confusable. Phonological similarity always penalized the good readers more than the poor ones. A further experiment (Shankweiler & I. Y. Liberman, 1976) showed that it made practically no difference whether the items to be recalled were presented to the eye or to the ear. These results strongly suggest that the difference between good and poor readers in recall of linguistic items will turn on their ability to use a phonological representation, whether derived from print or speech, and not merely on their ability to recode from print.

We might digress for a moment to ask whether the poor reader's problem may be a general deficit in short-term memory, or whether it is, indeed, a deficit specific to the processing of linguistic information. In a recent study directed to that question (I. Y. Liberman & Shankweiler, in press; I. Y. Liberman, Mark, & Shankweiler, 1978),[4] it was found that good and poor readers could not be distinguished on a recognition memory task employing photographed faces and abstract nonsense figures, but did differ significantly in their memory for nonsense syllables. This finding, and other existing evidence (see Vellutino, 1977, for a review), is consistent with the conclusion that the deficiencies of poor readers on memory tasks are limited to situations in which phonological representation can readily occur, either because the stimuli are linguistic items to begin with, or because they are objects to which verbal labels can readily be applied.

Returning, now, to the principal point, we should note that our original findings with letters apply to other linguistic materials and to other kinds of tasks closer to real reading situations. Two experiments speak to this matter. The first (Mark, Shankweiler, I. Y. Liberman, & Fowler, 1977) used (rhyming and nonrhyming) words instead of letters. It also had the advantage over the earlier study of a procedure that eliminated the possibility of differential rehearsal effects. Once again, the superior readers were much more strongly penalized by the confusable items than the poor readers.

In the second and more recent experiment, we have moved on to sentences. For this experiment (Mann, I. Y. Liberman, Shankweiler, & Katz, in preparation), we tested good and poor readers in recall of meaningful and semantically anomalous sentences, making a parallel comparison between conditions that did and did not offer the opportunity for phonological confusions to occur. A clear result of these new findings is that in recall of sentences, as with letters and words, good readers are much more affected than poor readers by phonological similarity.

There is, then, considerable support for the assertion that, for purposes of storing linguistic information in working memory, poor readers do not rely as much on a phonological strategy as good readers do. Given the effectiveness of the phonological strategy, and given that reading may put working memory under stress, especially in the beginner, we see that failure to use the phonology properly may be a cause, as well as a correlate, of poor reading.

As suggested in the Introduction, we suppose that the advantages of using phonological structures for short-term storage are independent of orthography and language. On that supposition, and given our results, we should anticipate that greater and lesser reliance on such structures might prove to be an important difference between good and poor readers everywhere.

## REFERENCES

Baddeley, A. D. How does acoustic similarity influence short-term memory? Quarterly Journal of Experimental Psychology, 1968, 20, 249-264.

Baddeley, A. D. Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. Quarterly Journal of Experimental Psychology, 1966, 18, 362-365.

Baddeley, A. D. Working memory and reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), The proceedings of the conference on visible language.

Eindhoven, in press.

Berko, J. The child's learning of English morphology. *Word*, 1958, *14*, 150-177.

Calfee, R., Chapman, R., & Venezky, R. How a child needs to think to learn to read. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

Chomsky, N. Comments for Project Literacy meeting. Project Literacy Report No. 2, pp. 1-8. Reprinted in M. Lester (Ed.), *Readings in applied transformational grammar*. New York: Holt, Rinehart, and Winston, 1970.

Chomsky, N., & Halle, M. *The sound pattern of English*. New York: Harper and Row, 1968.

Conrad, R. Acoustic confusions in immediate memory. *British Journal of Psychology*, 1964, *55*, 75-84.

Conrad, R. Speech and reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Downing, J. *Comparative reading*. New York: Macmillan, 1973.

Elkonin, D. B. USSR. In J. Downing (Ed.), *Comparative reading*. New York: Macmillan, 1973.

Erickson, D., Mattingly, I. G., & Turvey, M. T. Phonetic activity in reading: An experiment with kanji. *Haskins Laboratories Status Report on Speech Research*, 1973, *SR-33*, 137-156.

Fowler, C. A., Liberman, I. Y., & Shankweiler, D. On interpreting the error pattern of the beginning reader. *Language and Speech*, 1977, *20*, 162-173.

Fowler, C. A., Shankweiler, D., & Liberman, I. Y. Apprehending spelling patterns for vowels: A developmental study. *Haskins Laboratories Status Report on Speech Research*, in press, *SR-57*.

Gelb, I. J. *A study of writing* (Rev. ed.). Chicago: University of Chicago Press, 1963.

Gibson, E. J., & Levin, H. *The psychology of reading*. Cambridge, Mass.: MIT Press, 1975.

Gleitman, L. R., & Rozin, P. The structure and acquisition of reading I: Relations between orthographies and the structure of language. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading: The proceedings of the CUNY conference*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Harris, J. W. *Spanish phonology*. Cambridge, Mass.: MIT Press, 1969.

Helfgott, J. Phoneme segmentation and blending skills of kindergarten children: Implications for beginning reading acquisition. *Contemporary Educational Psychology*, 1976, *1*, 157-169.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, I. Y. Basic research in speech and lateralization of language: Some implications for reading disability. *Bulletin of the Orton Society*, 1971, *21*, 71-87.

Liberman, I. Y. Segmentation of the spoken word and reading acquisition. *Bulletin of the Orton Society*, 1973, *23*, 65-77.

Liberman, I. Y., Mark, L. S., & Shankweiler, D. Reading disability: Methodological problems in information-processing analysis. Letter to the editor, *Science*, 1978, *200*(4343), 801-802.

Liberman, I. Y., & Shankweiler, D. Speech, the alphabet, and teaching to read. In L. Resnick & P. Weaver (Eds.), *Theory and practice of early reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates, in press.

Liberman, I. Y., Shankweiler, D., Camp, L., Heifetz, B., & Werfelman, M. Steps toward literacy. A report prepared for the *Working Group on Learning Failure and Unused Learning Potential*, President's Commission on Mental Health, Washington, D.C., 1977.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 1974, *18*, 201-212.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, F. W. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. Scarborough (Eds.), *Toward a psychology of reading: The proceedings of the CUNY conference*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Mark, L. S., Shankweiler, D., Liberman, I. Y., & Fowler, C. A. Phonetic recoding and reading difficulty in beginning readers. *Memory & Cognition*, 1977, *5*, 623-629.

Martin, S. E. Nonalphabetic writing systems. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Mattingly, I. G. Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. Does awareness of speech as a sequence of phones arise spontaneously? Mimeo draft, Free University of Brussels, 1978.

Moskowitz, B. A. On the status of vowel shift in English. In T. Moore (Ed.), *Cognitive development and acquisition of language*. New York: Academic Press, 1973.

Read, C. Children's categorizations of speech sounds in English. NCTE research report 17, ERIC, 1975.

Rosner, J., & Simon, D. P. The auditory analysis test: An initial report. *Journal of Learning Disabilities*, 1971, *4*, 40-48.

Rozin, P., & Gleitman, L. R. The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading: The proceedings of the CUNY conference*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Savin, H. B. What the child knows about speech when he starts to learn to read. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Shankweiler, D., & Liberman, I. Y. Misreading: A search for causes. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Shankweiler, D., & Liberman, I. Y. Exploring the relations between reading and speech. In R. M. Knights & D. J. Bakker (Eds.), *Neuropsychology of learning disorders: Theoretical approaches*. Baltimore: University Park Press, 1976.

Treiman, R. A. Children's ability to segment speech into syllables and phonemes as related to their reading ability. Unpublished manuscript, Department of Psychology, Yale University, 1976.

Tzeng, O. J. L., Hung, D. L., & Wang, W. S.-Y. Speech recoding in reading

Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 621-630.

Vellutino, F. Alternative conceptualizations of dyslexia: Evidence in support of a verbal-deficit hypothesis. *Harvard Educational Review*, 1977, *47*, 334-354.

Zifcak, M. Phonological awareness and reading acquisition in first grade children. Unpublished doctoral dissertation, University of Connecticut, 1977.

## Footnotes

[1] Chomsky refers to this form as the "lexical representation." But since we wish to consider later whether this or some other representation is the actual basis of lexical lookup, and so deserves to be called *the* lexical representation, we use the neutral and descriptive term "morphophonological representation" instead.

[2] This circumstance presents a hazard for anyone who has to teach beginners to read an alphabetic orthography. Given the impossibility of producing many consonants in isolation, how does the teacher help the child to identify the linguistic units that the orthography represents? If the teacher "sounds out" the consonants by coarticulating them with the neutral vowel [ə], a very common strategy, she runs the risk of confusing the child, for surely the syllable that results is inappropriate for almost all of the contexts in which the consonant will be represented in printed text. Possible ways around this difficulty have been discussed in detail elsewhere (I. Y. Liberman et al., in press; I. Y. Liberman, Shankweiler, Camp, Heifetz, & Werfelman, 1977).

[3] We should note that other investigators besides ourselves have remarked on the difficulty of becoming aware of the phonemic segment, and also on the possibility that this might be a problem in learning to read. Among these are Calfee, Chapman,& Venezky, 1972; Downing, 1973; Elkonin, 1973; Gibson & Levin, 1975; Gleitman & Rozin, 1977; Rosner & Simon, 1971; Rozin & Gleitman, 1977; Savin, 1972; Vellutino, 1977.

[4] A full account of this study, which includes M. Werfelman as a co-author, is in preparation.

# ASPIRATION AMPLITUDE AS A VOICING CUE FOR SYLLABLE-INITIAL STOP CONSONANTS PRESENTED MONAURALLY AND IN DICHOTIC COMPETITION[*]

Bruno H. Repp

Abstract. The present experiments demonstrate that amplitude of aspiration noise (relative to the following vocalic portion) is a cue for the distinction between voiced and voiceless syllable-initial stop consonants in English, and that it can be traded for voice onset time (VOT). In Experiment I, the category boundary on a synthetic VOT continuum, /da/-/ta/, was found to be a linear function of the amplitude ratio between the aspirated and vocalic portions over a 24-dB range. In Experiment II, the synthetic stimuli were prefixed with a natural 10-msec release burst, and burst and aspiration amplitudes were varied orthogonally. Both factors affected the voicing boundary in the expected direction but not independently; their interaction was ascribed to backward masking of weak bursts by strong aspiration noise. After accounting for this interaction, the effect of burst amplitude seemed very small compared to that of aspiration amplitude. These results suggest that the amount of aspiration noise is the primary voicing cue, and that the perception of the noise follows psychoacoustic laws of time-intensity tradeoff. Experiment III used the findings of Experiment I to test the "category goodness" hypothesis of dichotic competition, which predicted that increasing the aspiration amplitude of a /ta/ syllable should increase the perceptual dominance of this stimulus over a /da/ simultaneously presented to the other ear. This hypothesis was not supported. The results indicated that dichotic stimulus dominance was determined by psychoacoustic factors, not by phonetic category goodness.

## INTRODUCTION

Many studies have investigated the acoustic cues for various phonetic distinctions, such as voicing or place of articulation. A number of different perceptual cues is known to exist for each such distinction; for example, both voice onset time (VOT) and the onset frequency of the first formant are important voicing cues in syllable-initial stop consonants (Lisker, 1975;

[*] A version of this paper including only Experiments I and II will be published in Language and Speech.

Acknowledgment: This research was supported by NICHD Grant HD01994 and BRS Grant RR05596. I would like to thank Patti Price for running Experiment II, Arthur Abramson, Leigh Lisker, and especially Virginia Mann for stimulating discussions, and Alvin Liberman, Dominic Massaro, and especially Arthur Abramson for comments on an earlier draft.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

Summerfield & Haggard, 1977). Although the acoustic stimulus properties we call cues are generally not independent in articulation, they can be independently manipulated in perceptual experiments, and they are often found to make independent contributions to a given perceptual distinction.

Several recent experiments have investigated trading relations between different cues for the same phonetic contrast: A change in one cue can be compensated for, within limits, by an opposing change in another cue, so that exactly the same phonetic percept results (Bailey & Summerfield, 1978; Massaro & Cohen, 1976, 1977; Repp, Liberman, Eccardt, & Pesetsky, 1978; Summerfield & Haggard, 1977). Such trading relations are a natural consequence of the multiplicity of cues for phonetic contrasts. More importantly, perhaps, these experiments have provided information about the relative importance of different cues in signaling a given phonetic distinction.

In part, the perceptual salience of a given cue may be a direct consequence of its qualitative auditory properties; e.g., for a given phonetic distinction, there may be a natural hierarchy of cues, such that spectral cues are more important than temporal cues, or some spectral cues are more important than others. However, the perceptual weight of a cue also depends on its prominence or clarity relative to the other parts of the signal. Among the quantitative auditory parameters that determine relative prominence are amplitude and formant bandwidth. These "secondary" parameters have received much less attention in past investigations than the "primary" parameters of duration and spectral frequency. Nevertheless, the former are likely to play an important role in perception. For example, it is almost certain--though rarely pointed out in the literature--that the relative amplitudes of the second and third formants determine the relative weights of their respective transitions as cues for place of articulation (cf. Repp, 1978b). By synthesizing speech or by manipulating real-speech tokens, we often disturb the natural cue hierarchy and may be misled as to the relative importance of different cues; therefore, we should pay close attention to factors determining perceptual prominence.

The present experiments examined whether the amplitude relationship between the aspiration noise and the following vocalic portion constitutes a cue for the perception of the voicing distinction in syllable-initial stop consonants. Voicing contrasts in English are conveyed by a variety of acoustic cues, all of which have been postulated to be consequences of an underlying change in laryngeal timing (Lisker & Abramson, 1964, 1971). Not surprisingly, the temporal aspect of this cue complex has been found to be most prominent perceptually, and many studies have varied the delay between stop release and voicing onset (i.e., VOT)[1] in synthetic syllables to determine the perceptual boundary between the voiced and voiceless categories. Emphasis on the timing aspect has led to a relative neglect of the fact that the interval prior to voicing onset is filled with aspiration noise, whose very presence is likely to constitute a cue for voicelessness. This must be particularly true in English, where--as every linguist knows--the phonological categories "voiced" and "voiceless" actually represent the phonetic distinction between voiceless unaspirated (occasionally voiced) and voiceless aspirated stops.

Thus, it may be argued that the primary perceptual cue is not the abstract temporal property of "delay in voicing onset" (called "separation" by Summerfield & Haggard, 1974), but the presence and amount of aspiration during that delay, even though these two properties are tightly correlated in natural speech. We might expect, then, that an increase in the amplitude of the aspiration noise relative to the following vocalic portion would increase the salience of this important cue and, thus, increase the probability of classifying a syllable-initial stop consonant in the voiceless (aspirated) category.

Two earlier studies have given some attention to the perceptual effects of aspiration in syllable-initial stops. Winitz, LaRiviere, and Herriman (1975) actually varied aspiration intensity in one of their experiments, but their results were not clear-cut, due in part to ceiling effects. Summerfield and Haggard (1974) found that presence vs. absence of aspiration noise following the release burst had little (or even a paradoxical) effect on voicing perception. However, they did not report the amplitude of their noise source, which may have been relatively weak.

Experiment I was conducted to investigate the perceptual trading relation between relative amplitude and duration of aspiration (i.e., VOT) as joint voicing cues. The design of the experiment also permitted a second question to be asked, viz., whether changes in overall (absolute) stimulus amplitude within a 12-dB range affect voicing perception.

## EXPERIMENT I

### Method

*Subjects*. Eight subjects participated. They included the author, a research assistant, and six paid volunteers (Yale undergraduates) who had participated in previous experiments using synthetic syllables and had proven to be reliable listeners.

*Stimuli*. The stimuli were generated with the OVE IIIc serial resonance synthesizer at Haskins Laboratories. All stimuli were stop-consonant-vowel syllables perceived as either /da/ or /ta/. Their total duration was 300 msec. Fundamental frequency was constant at 125 Hz over the first 84 msec and then fell linearly to 90 Hz at offset. The initial formant transitions were stepwise-linear and 48 msec in duration; $F_1$ rose from 285 to 771 Hz, $F_2$ fell from 1543 to 1233 Hz, and $F_3$ fell from 3019 to 2520 Hz. The duration of the synthesis time frames was 4 msec.

A 10-member VOT continuum was created by replacing periodic excitation with noise and simultaneously increasing the bandwidth of $F_1$ to its maximum (thereby essentially eliminating $F_1$). The amplitude of the aspirated portion was about 20 dB below that of the following vocalic (periodic) portion, as determined by measurement of the synthesizer output. The periodic source was turned on 8 msec (one pitch period) before voicing onset but kept at a minimal amplitude. This procedure insured that the second pitch pulse, which marked the true onset of voicing, had full amplitude. The ten stimuli thus generated had VOTs ranging from 8 to 44 msec in 4-msec steps. They had no special release bursts at onset.

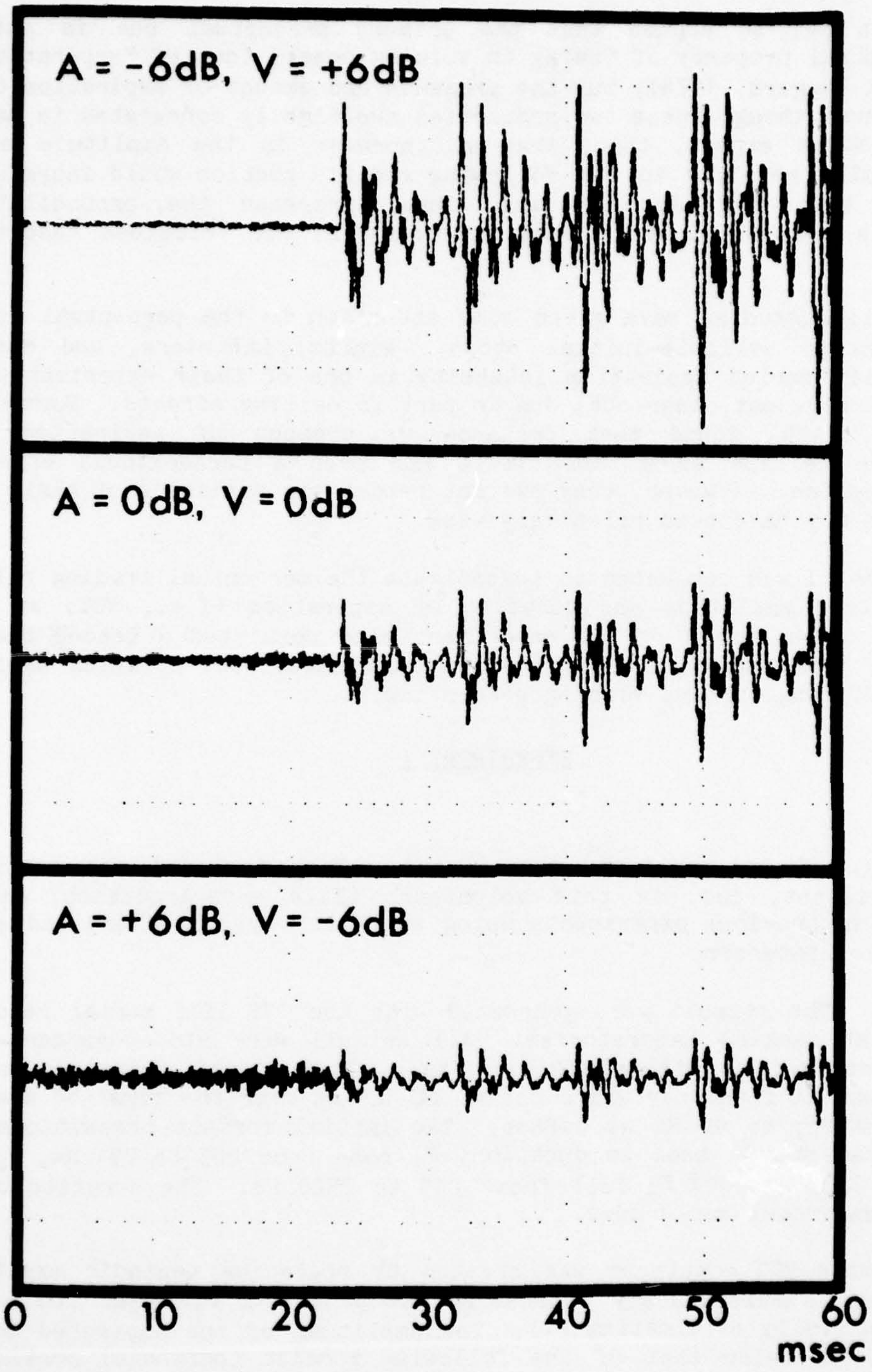


Figure 1: Oscillograms of the first 60 msec of a representative stimulus (VOT = 24 msec) in three conditions of Experiment I.

All stimuli were digitized at 10 kHz using the Haskins Laboratories pulse code modulation system. From the digitized waveforms, eight additional stimulus series were constructed by independently amplifying or attenuating the aspirated and vocalic portions of the original stimuli. Changes in amplitude in either stimulus portion were achieved by means of a computer instruction after placing a cursor at the onset of the first true pitch pulse. These manipulations resulted in a total of nine stimulus series, each a 10-member VOT continuum. Relative amplitudes of the aspirated portion of either -6, 0, or +6 dB (relative to the original stimuli) were orthogonally combined with relative amplitudes of the vocalic portion of either -6, 0, or +6 dB (relative to the original stimuli). Thus, the stimulus ensemble included both a 12-dB range in absolute stimulus amplitude (from the -6/-6 to the +6/+6 stimulus series; the slash symbolizes the partition into aspirated and vocalic stimulus portions), and a 24-dB range in the relative amplitudes of aspirated and vocalic portions (from the -6/+6 series at one extreme to the +6/-6 series at the other). This range is illustrated in Figure 1 which shows oscillograms of the first 60 msec of a representative stimulus (VOT = 24 msec) in the -6/+6, 0/0, and +6/-6 conditions. Given a true amplitude ratio between periodic and nonperiodic portions of about 20 dB in the 0/0 condition, the total range extended from 8 dB (+6/-6 condition) to 32 dB (-6/+6 condition).

Each of the nine stimulus series yielded a basic test unit of 28 stimuli, since the 10 stimuli in each series were recorded with the following frequencies: 1,2,3,4,4,4,4,3,2,1. Thus, four times as many responses were collected for the center stimuli (which were likely to bracket the voicing boundary) than for the endpoint stimuli (which were likely to be reliably classified as voiced and voiceless, respectively). The complete stimulus sequence contained 9 x 28 = 252 stimuli in random order, with interstimulus intervals of 2 sec. Two such sequences were recorded. Some practice stimuli (endpoint stimuli at different overall intensities) preceded the first series.

Procedure. Each subject participated in two sessions, the second session being an exact replication of the first. The task was to identify in writing each syllable as beginning with either a D or a T. All in all, each subject listened to 4 blocks of 252 stimuli, providing a total of 16 responses to each stimulus with one of the four critical VOTs (20-32 msec) in the voicing boundary region.

The tape was played back on an Ampex AG-500 tape recorder, and the subjects listened binaurally over Telephonics TDH-39 earphones. The amplitude was calibrated by means of a series of rapidly repeated /da/ syllables; the peak amplitude for this calibration series, registered by a Hewlett-Packard volt meter, was adjusted to a level of about 80 dB SPL, which approximates the absolute amplitude of the vocalic portion. The lowest intensity of the aspiration noise in the course of the experiment was about 54 dB SPL and thus well above the (unmasked) detection threshold.

## Results

The results are graphically displayed in Figure 2. The three panels show the effect of varying the amplitude of the aspiration noise (A) at each of the three levels of the amplitude of the vocalic portion (V). Comparing the response functions within each panel, it can be seen that the number of D

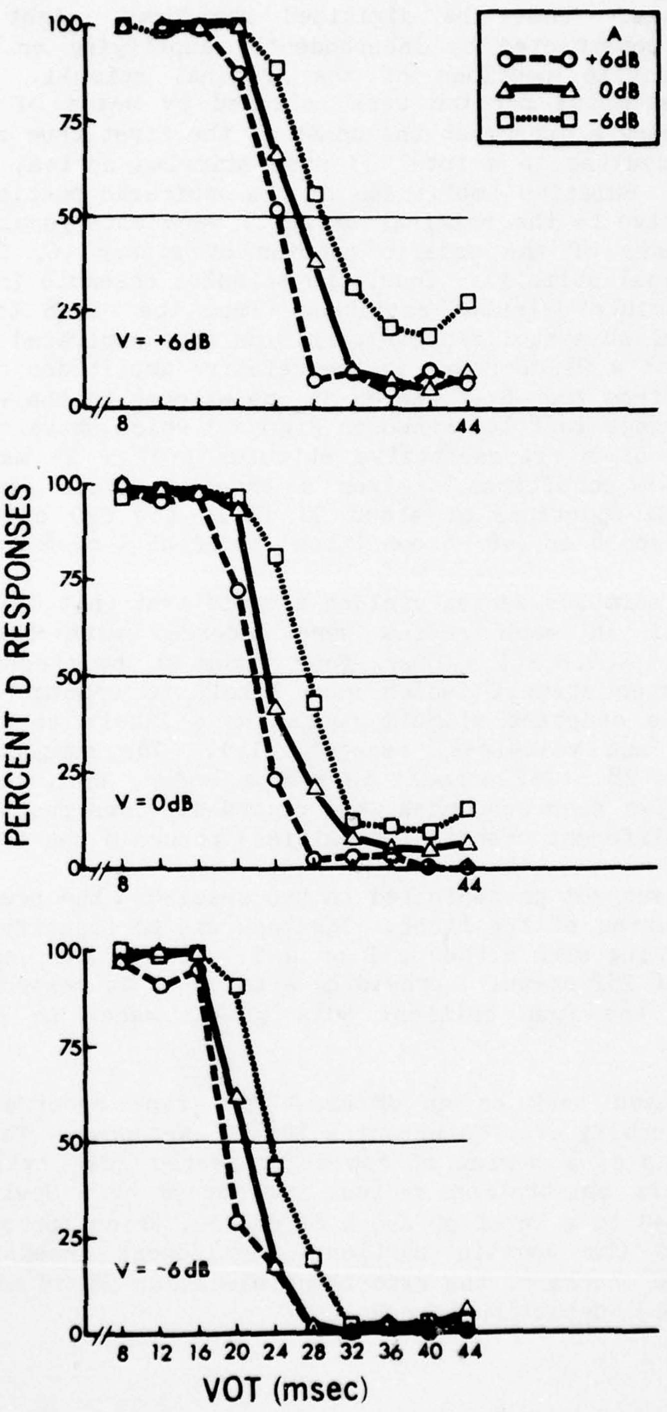
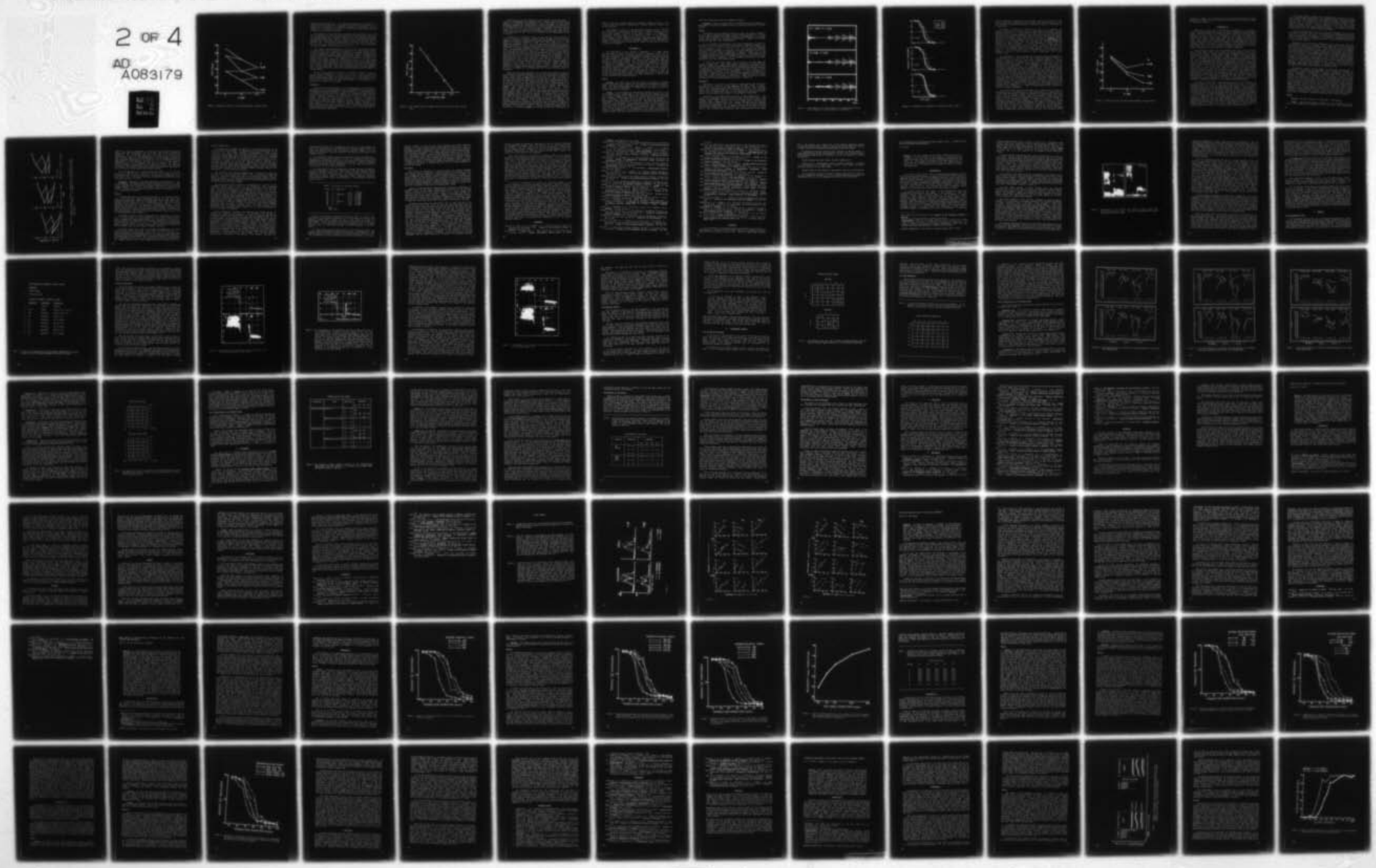
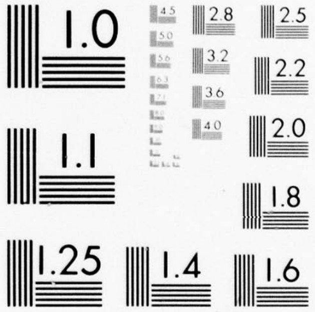


Figure 2: Percentage of D responses as a function of VOT, A, and V.

AD-A083 179    HASKINS LABS INC  NEW HAVEN CONN    F/G 5/10
SPEECH RESEARCH, A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
MAR 79  A S ABRAMSON, T BAER, C BEST, G CARDEN  PHS-HD-01994
UNCLASSIFIED  SR-57(1979)    NL

2 OF 4
AD
A083179





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

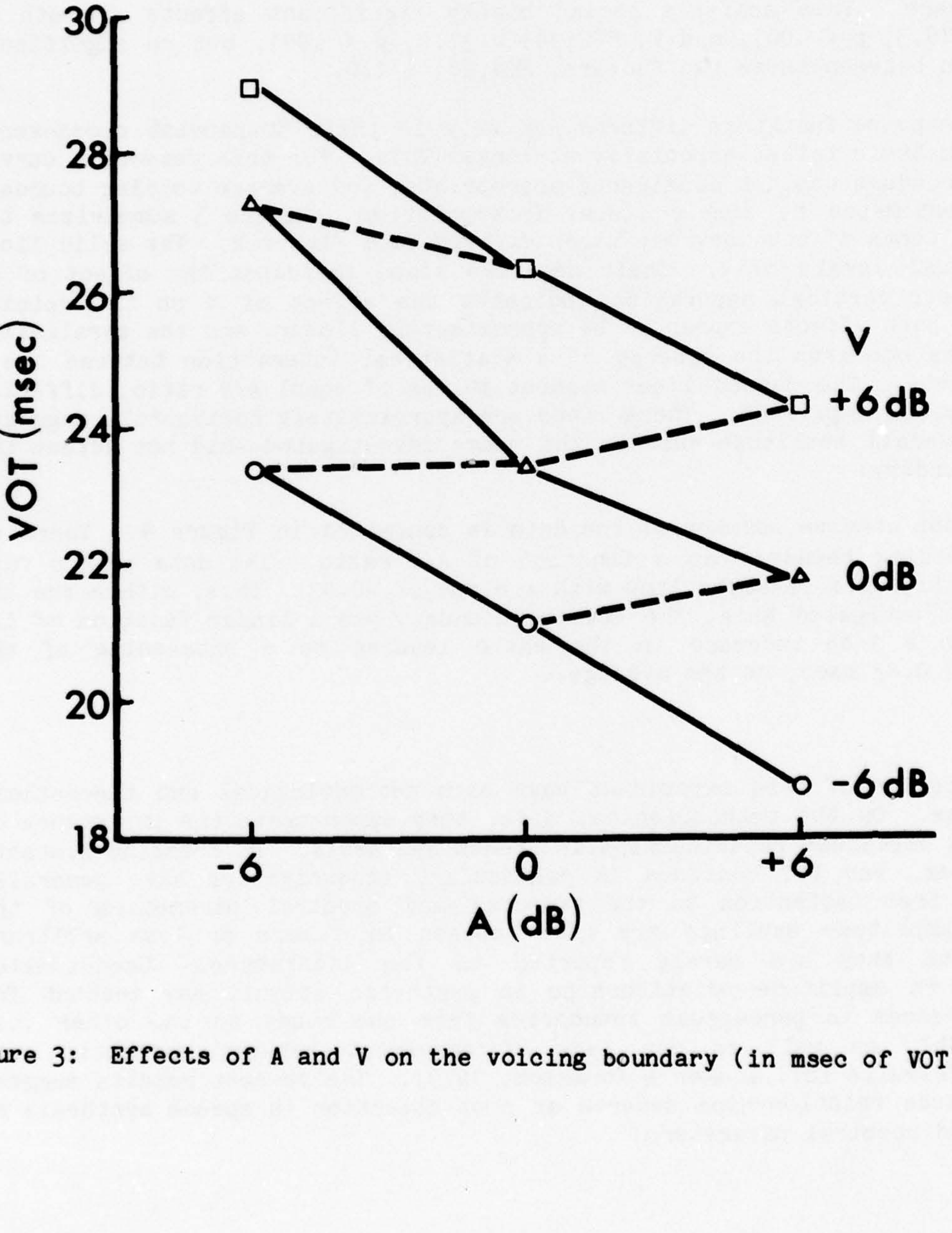


Figure 3: Effects of A and V on the voicing boundary (in msec of VOT).

responses decreased (and that of T responses increased) as A increased, in accord with the predictions. Comparing the functions across panels, it is evident that the number of D responses decreased as V decreased. Since a decrease in V increased A relative to V (the A/V ratio), this result was equally in line with the predictions.

Two separate two-way analyses of variance were conducted, one on the total frequencies of voiced (D) responses in each stimulus series, and one on the voicing boundaries (50-percent intercepts of the labeling functions). Both analyses gave similar results, and since some of the individual boundary estimates were unreliable, only the results of the first analysis will be reported here. This analysis showed highly significant effects of both A, $F(2,14) = 78.3$, $\underline{p} < .001$, and V, $F(2,14) = 37.6$, $\underline{p} < .001$, but no significant interaction between these two factors, $F(4,28) = 1.0$.

The response functions differed not only in their 50-percent crossovers, but also in their tails, especially at longer VOTs. For this reason, a curve-fitting procedure was not considered appropriate, and average voicing boundaries were estimated by simple linear interpolation. Figure 3 summarizes the results in terms of boundary estimates derived from Figure 2. The solid lines connect equal levels of V. Their negative slope indicates the effect of A, whereas their vertical separation indicates the effect of V on the voicing boundary. Both effects appear to be approximately linear, and the parallelism of the lines confirms the absence of a statistical interaction between the A and V effects. The dashed lines connect points of equal A/V ratio, differing only in overall amplitude. These lines are approximately horizontal, suggesting that overall amplitude--within the range investigated--did not affect the voicing boundary.

The most concise summary of the data is contained in Figure 4. There we see the voicing boundary as a function of A/V ratio. The data points fall almost exactly on a straight line with a slope of -0.43. Thus, within the 24-dB range investigated here, the voicing boundary was a linear function of A/V ratio, with a 1-dB increase in the ratio leading to a shortening of the boundary by 0.43 msec, on the average.

## Discussion

The results of this experiment have both methodological and theoretical implications. On the methodological side, they demonstrate the importance of considering amplitude relationships in speech synthesis. In creating synthetic syllables, and VOT continua in particular, investigators have generally restricted their attention to the temporal and spectral parameters of the stimuli. Amplitude settings are often chosen in a more or less arbitrary fashion, and they are rarely reported in the literature. Uncontrolled variations in amplitude relationships in synthetic stimuli may account for some differences in perceptual boundaries from one study to the other (cf. Repp, 1978b), as well as for lack of agreement between perception and production results (cf. Lisker & Abramson, 1970). The present results suggest that amplitude relationships deserve as much attention in speech synthesis as temporal and spectral parameters.

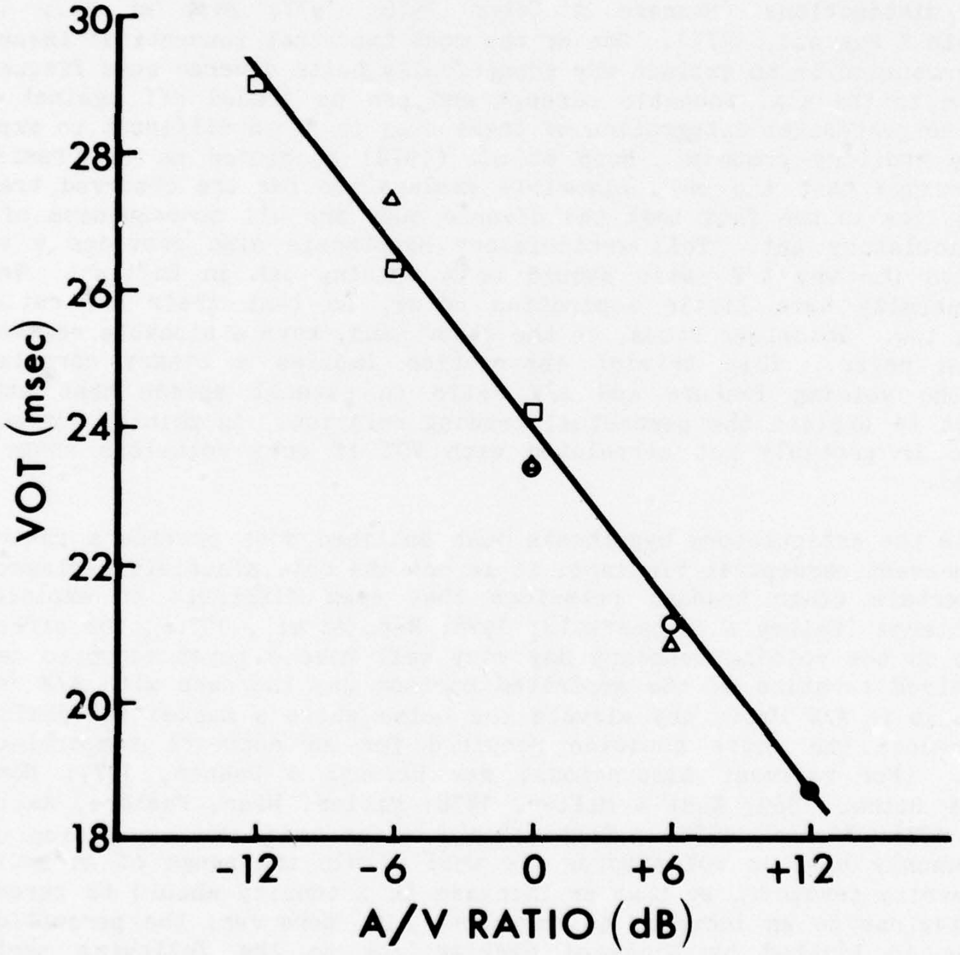


Figure 4: The trading relation between VOT and A/V ratio as joint voicing cues.

Amplitude has also been neglected in the analysis of natural speech. A number of investigators have measured VOTs in natural speech without paying attention to amplitude relationships.[2] Current knowledge of articulatory mechanisms suggests that speakers should have little control over A/V ratio.[3] Any increases in A due to an increase in air flow probably would also increase V and thus leave the A/V ratio unchanged. Therefore, A/V ratio may be expected to vary only randomly across productions of voiceless stops and not to be correlated with VOT. Still, the relevant acoustic measurements remain to be done.

Turning now to theoretical issues, we note that the linear function in Figure 4 represents a perceptual trading relation between A/V ratio and VOT, comparable to similar trading relations found between multiple cues for other phonetic distinctions (Massaro & Cohen 1976, 1977; Repp et al., 1978; Summerfield & Haggard, 1977). One of the most important theoretical issues in speech perception is to explain why acoustically quite diverse cues frequently contribute to the same phonetic percept and can be traded off against each other. The perceptual integration of these cues is often difficult to explain on purely auditory grounds. Repp et al. (1978) concluded on the basis of their findings that the only plausible explanation for the observed trading relations lies in the fact that the diverse cues are all consequences of the same articulatory act. This articulatory hypothesis also provides a valid explanation for why A/V ratio should be a voicing cue in English. Voiced stops generally have little aspiration noise, so that their A/V ratio is naturally low. Voiceless stops, on the other hand, have a sizeable segment of aspiration noise. This trivial observation implies a binary correlation between the voicing feature and A/V ratio in natural speech that may be sufficient to explain the perceptual trading relation. As pointed out above, A/V ratio is probably not correlated with VOT if only voiceless stops are considered.

While the articulatory hypothesis just outlined does provide a rationale for the present perceptual findings, it is not the only plausible explanation. Unlike certain other trading relations that seem difficult to explain in auditory terms (Bailey & Summerfield, 1978; Repp et al., 1978), the effect of A/V ratio on the voicing boundary may very well have a psychoacoustic basis: The perceived duration of the aspirated portion may increase with A/V ratio, the increase in A/V ratio may elevate the noise above a masked threshold, or it may reduce the noise duration required for an accurate temporal-order judgment. (For relevant discussions, see Divenyi & Danner, 1977; Homick, Elfner, & Bothe, 1969; Kuhl & Miller, 1978; Miller, Wier, Pastore, Kelly, & Dooling, 1976; Pisoni, 1977). Pastore[4] has noted that the aspiration durations commonly used in VOT studies lie well within the range of an auditory time-intensity tradeoff, so that an increase in intensity should be perceptually equivalent to an increase in duration. If, moreover, the perception of the noise is limited by backward masking due to the following periodic portion, any time-intensity tradeoff would presumably be enhanced. The occurrence of such a tradeoff in the perception of VOT would undermine one form of psychoacoustic explanation according to which the abstract property of temporal separation (between stimulus onset and voicing onset, as measured in the signal) is the major voicing cue. Rather, the tradeoff suggests that the critical cue is aspiration _energy_; the voicing boundary then represents the point at which this energy exceeds the listener's response criterion. This criterion may (but need not) coincide with a masked detection threshold for the noise portion.

This argument in favor of a psychoacoustic mechanism does not diminish the plausibility of articulatory explanations for tradeoffs between acoustically quite different portions of the speech signal. For example, an articulatory rationale seems to be required to account for the perceptual integration of $F_1$ onset frequency and VOT as joint cues to voicing (Summerfield & Haggard, 1977). However, the present tradeoff between amplitude and duration of aspiration noise occurs between two aspects of the same acoustic segment, and such trading relations are more likely to take place at the level of auditory processing.

## EXPERIMENT II

The significance of the results of Experiment I to normal speech perception must be qualified by the fact that the synthetic stimuli did not have any release bursts. Burstless synthetic stimuli have been used in numerous experiments, and the methodological conclusions reached above apply to these studies in particular. However, it is possible that, when the release is more clearly marked by a plosive burst, aspiration amplitude decreases in perceptual salience. This is especially suggested by the study of Summerfield and Haggard (1974) who found no perceptual effect of presence vs. absence of aspiration following a release burst (in /-a/ context). Perhaps the variations in aspiration amplitude in burstless stimuli serve only to mark the moment of stimulus onset more or less clearly. In order to investigate this hypothesis, a second experiment was conducted in which the stimuli had an initial release burst. The amplitudes of that noise burst and of the following aspiration noise were varied orthogonally to reveal their respective effects on voicing perception.

### Method

Subjects. The author and seven new subjects (including one graduate research assistant, one colleague, and five paid student volunteers) participated. The paid subjects were less experienced in listening to synthetic speech than those in Experiment I; however, this was made up for by the improved quality of the stimuli.

Stimuli. A new 10-member VOT continuum was synthesized, similar to the basic stimulus series in Experiment I (0/0 condition) but ranging in VOT from 0 to 36 msec in 4-msec steps. These stimuli were digitized and then prefixed with a 10-msec alveolar release burst, so that all VOTs were increased by 10 msec (10-46 msec). The burst was taken from the digitized waveform of a natural-speech /da/ pronounced by the author. Using the same methods as in Experiment I, the amplitude of the 10-msec burst portion (B) and the amplitude of the following, synthetic aspirated portion (A) were varied orthogonally in three steps (-6, 0, +6 dB relative to the baseline stimuli), leading to a total of nine stimulus series. The amplitude of the vocalic portion was not varied in this experiment. The baseline (0/0, the slash now denoting the division between burst and aspiration) and the extreme conditions (+6/-6 and

-6/+6) are illustrated by the oscillograms in Figure 5.

Procedure. Design, stimulus tapes, and procedure were all analogous to Experiment I, with the sole exception that all data were collected in a single session.

## Results

The effects of the two amplitude factors, B and A, are shown in Figure 6 which is exactly analogous to Figure 2, with B taking the place of V. It can be seen that, in general, there was a decrease in D responses as A increased (within panels) and as B increased (across panels). However, the effects seemed less systematic than in Experiment I.

A 3 x 3 analysis of variance of the total frequencies of voiced (D) responses in the nine stimulus series yielded highly significant effects of B, F(2,14) = 14.0, $p < .001$, of A, F(2,14) = 32.4, $p < .001$, and a significant B x A interaction, F(4,28) = 19.3, $p < .001$. The results were extremely consistent across subjects, as indicated by the high significance levels. The interaction is more clearly represented in Figure 7 which plots voicing boundaries derived by linear interpolation from Figure 6. Surprisingly, B had its largest effect when A was at its highest level; when A was low, B had only a negligible effect. Similarly, A had its strongest effect when B was high; when B was low, A had no systematic effect.

It should be noted that, despite the interaction, the noise amplitude effect of Experiment I was almost exactly replicated: Considering only amplitude changes in the total noise portion preceding voicing onset (-6/-6, 0/0, and +6/+6 conditions), we find that the corresponding voicing boundaries in Experiment II fell approximately on a straight line with a slope of -0.50, as compared with a slope of -0.43 in Experiment I. It is also evident from a comparison of Figures 3 and 7 that the voicing boundaries in Experiment II were about 5 msec longer than in Experiment I, suggesting that the 10-msec burst was perceptually equivalent to only about 5 msec of aspiration noise, perhaps due to its nonuniform amplitude contour (cf. Figure 5).

## Discussion

Experiment II fulfilled its purpose by demonstrating that aspiration amplitude has an effect on voicing judgments even when a burst is present, particularly (and paradoxically) when the burst is strong. Thus, the hypothesis that aspiration amplitude in burstless stimuli has its perceptual effect solely by marking the onset of the stimulus can be safely rejected. Rather, the total noise portion seems to be perceptually significant.

Experiment II also demonstrated an effect of burst amplitude (B). This effect is in agreement with the observation that voiced stops tend to have somewhat weaker bursts (relative to the following vocalic portion) than voiceless stops in English (Zue, 1976). Thus, there may be an articulatory basis for the perceptual effect of B. However, we must seriously consider psychoacoustic explanations for the present results, particularly to explain the curious finding that B had an effect only when A was high (and vice versa). This finding seems paradoxical when considering what would happen

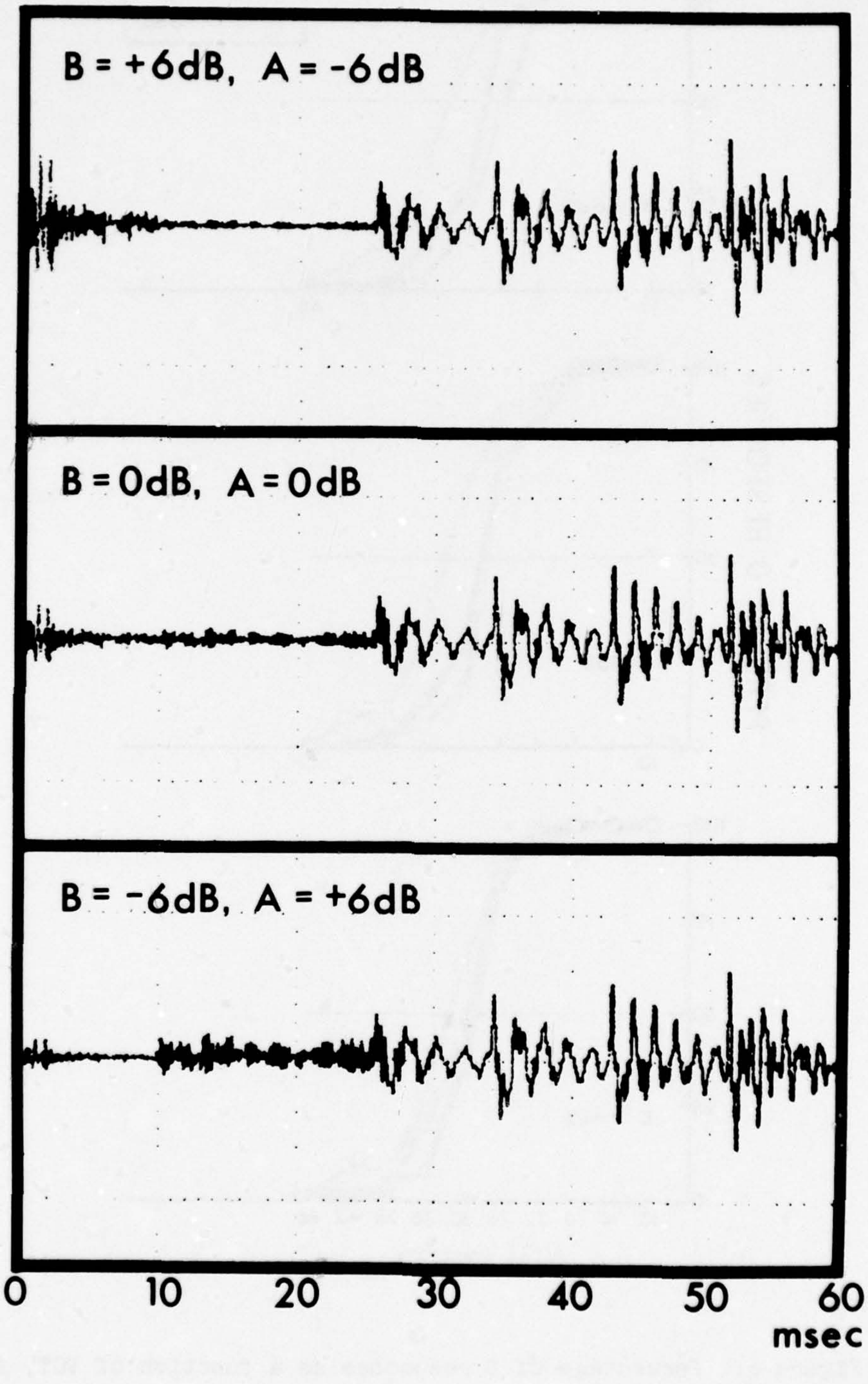


Figure 5: Oscillograms of the first 60 msec of a representative stimulus (VOT = 26 msec) in three conditions of Experiment II.

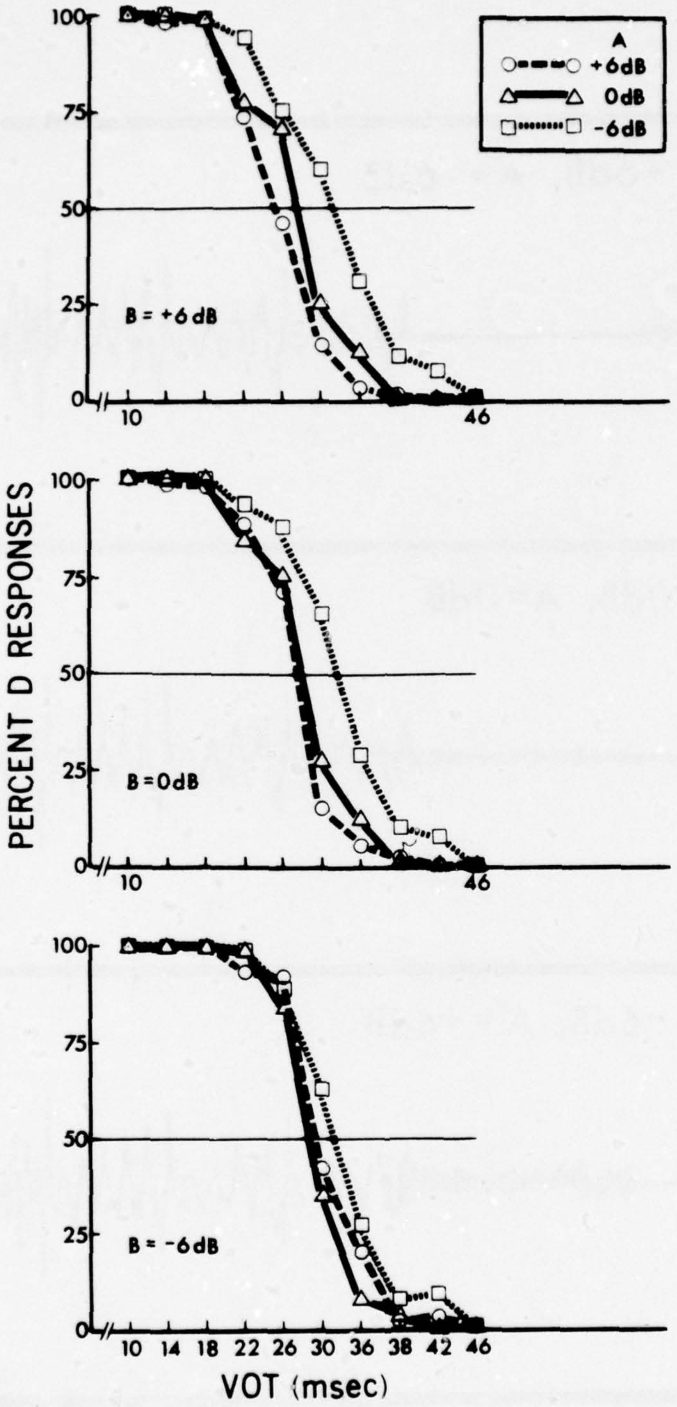


Figure 6: Percentage of D responses as a function of VOT, A, and B.

when B is gradually reduced to minus infinity. Since the effect of A was already small and nonsystematic at the lowest burst intensity used here (-6 dB), no effect of A would be predicted in burstless stimuli, in contradiction to Experiment I.

The puzzle is resolved as follows:[5] Note that if B is minimal, the original burst duration (10 msec) must be subtracted from the VOT. If B is gradually lowered, it is likely that the burst suffers increasing backward masking by the following aspiration noise, which reduces the effective burst duration (and thus the effective VOT) until no burst is perceived any longer. The results in Figure 7 have been plotted on a scale of physical VOT (as measured in the acoustic signal), but we do not know the psychological VOT perceived by the listener after the occurrence of auditory interactions between the signal components. It seems reasonable to assume that the psychological VOT was reduced in those stimuli in which a weak burst was followed by a strong aspiration noise, so that the data points for the -6/+6, -6/0, and 0/+6 conditions should be shifted downward on the VOT scale (cf. Figure 7). Such a shift would tend to eliminate the B x A interaction and generate a pattern of results similar to that of Experiment I, viz., three parallel lines.

One consequence of this interpretation is that the effect of B appears to be negligible once a correction for the hypothetical masking effect is applied. Varying B helps the burst evade backward masking by a strong aspiration noise, but it seems to have little direct cue value for the voiced-voiceless distinction in English.[6] What remains is a strong effect of A, in accordance with Experiment I. It is possible to give this effect an interpretation similar to that of B in Experiment II: The aspiration noise may suffer backward masking from the powerful periodic portion that follows, and changes in amplitude in either stimulus portion may change the effective temporal relation between them, perhaps by affecting neural transmission times to the centers of speech perception. One might predict that A would lose its cue value once it exceeds the values at which backward masking occurs, but since such values were presumably not reached in the present experiments, the consistent effects of A in all conditions do not contradict a backward masking interpretation. Thus, the hypothesis that the significant voicing cue is the abstract temporal relation between the onsets of nonperiodic and periodic portions may be resurrected if it is understood that this timing relation is a perceptual one, determined at a stage following psychoacoustic interactions between the signal components.

To conclude, the results of Experiments I and II should not be taken as a demonstration that VOT perception, or voicing perception, or perhaps even all of speech perception can and should be explained by psychoacoustic principles alone. They do suggest, however, that the alternative extreme view, that speech perception is to be understood entirely in terms of apprehending the behavior of an articulatory system, is similarly untenable. Instead of fostering another artificial dichotomy, it seems more reasonable to assume that speech perception involves mechanisms at a number of different levels. Psychoacoustic factors often seem prominent because of the psychoacoustic problems that we create for listeners by manipulating speech signals in certain arbitrary ways. Almost certainly, some of the auditory processes revealed in experiments such as the present ones play a role in natural speech

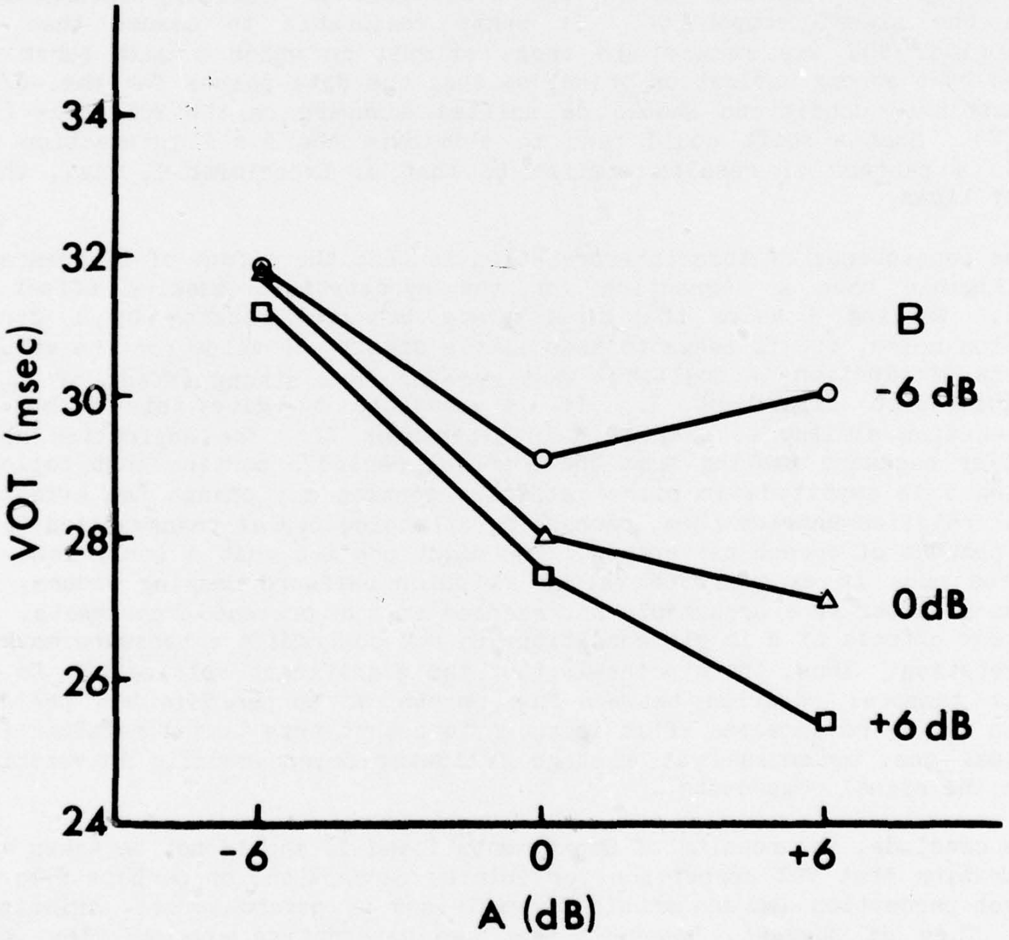


Figure 7: Effects of A and B on the voicing boundary (in msec of VOT).

perception. However, it is equally true that the guiding principle of speech perception at higher levels is likely to be found in the articulatory origin of the auditory signal.

## EXPERIMENT III

Experiment III made use of the results of Experiment I in testing a specific hypothesis about the nature of dichotic competition between speech sounds. When two phonetically conflicting speech sounds are presented simultaneously to the two ears, it is frequently the case that one perceptually dominates the other. In particular, when the two stimuli are fused, the single fused stimulus often sounds more like one component than like the other. Repp (1976a, 1977a) hypothesized that this phenomenon--the dichotic stimulus dominance effect--reflects the relative category goodness of the two competing stimuli. According to this hypothesis, that stimulus which corresponds more closely to the listener's perceptual prototype of a relevant phonetic category will be more dominant in dichotic competition.

Some preliminary support for this hypothesis has been found in experiments by Repp (1976a, 1976b, 1977a, 1978b), but strong evidence in favor of it has yet to be obtained. Indeed, several recent studies have failed to support the hypothesis (Pompino, Rilhac-Sutter, Simon, & Sommer, 1977; Repp, 1978c, 1978d). Experiment III was designed to provide a relatively straightforward test. It is known that the dichotic stimulus dominance relationship between two fused stop-consonant-vowel syllables contrasting in voicing (e.g., /da/-/ta/) can be substantially altered by changing the VOTs of the competing stimuli, particularly the VOT of the voiceless stimulus (Miller, 1977; Repp, 1977a, 1978a). A /ta/ with a long VOT dominates a /da/ in the other ear more than does a /ta/ with a short VOT. This is in agreement with the category goodness hypothesis, since the prototypical voiceless stop probably has a fairly long VOT. However, the effect of variations in VOT might also be explained by interactions at the auditory level. For example, the longer aspirated portion that goes with a longer VOT may simply be more detectable or more salient in a fused stimulus composed of /da/ and /ta/. Thus, the dependence of stimulus dominance on the VOTs of the competing stimuli cannot be taken as unequivocal support for the category goodness hypothesis.

The results of Experiment I suggested an alternative way of improving the category goodness of a /ta/, viz., increasing its A/V ratio. One way of achieving this is to lower the amplitude of the vocalic portion (V) while holding the amplitude of the aspirated portion (A) constant. In Experiment I, the resulting reduction in the overall amplitude of the syllable (which is determined primarily by the periodic portion) had no effect on the responses. In the context of a dichotic experiment, the procedure just mentioned creates a situation where the overall amplitude of /ta/ will be reduced relative to that of a competing /da/. If overall amplitude differences between the competing stimuli do not affect their dominance relationship (and there was some reason to believe that this might be the case--see below), the category goodness hypothesis predicts that a /ta/ with a higher A/V ratio will be more dominant over a /da/ than a /ta/ with a lower A/V ratio, even though its overall amplitude is lower. Let us call this the strong prediction.

Of course, there is another way of changing the A/V ratio of a /ta/, viz., by increasing the amplitude of its aspirated portion, holding everything else constant. This manipulation would slightly increase the overall amplitude of the /ta/ stimulus, and the expected resulting increase in its dominance over a /da/ would not be in conflict with (but most likely larger than) a possible effect on stimulus dominance of the overall amplitude relationship between the competing dichotic stimuli. This expectation was the weaker prediction of the category goodness hypothesis.

Thus, if the amplitudes of A and V are varied orthogonally in a /ta/ that is dichotically paired with a /da/, and if amplitude relationships across ears play no role (within the limits of the experiment), then the same results should be obtained as in Experiment I: There should be independent effects of A and V, and the subjects' responses should be a direct function of A/V ratio in the voiceless stimulus.

Experiment III also directly investigated the role that overall amplitude relationships between the competing stimuli might play in this particular situation. For this purpose, variations in the amplitude of the /da/ stimulus were included as an additional factor. Several earlier studies have examined the stimulus amplitude factor and found that it had relatively little effect. In unfused dichotic stimuli, changes in amplitude relationships seem to play a minor role as long as one amplitude does not get so low that the intelligibility of the stimulus in one ear is impaired (Cullen, Thompson, Hughes, Berlin, & Samson, 1974; Speaks & Bissonette, 1975). There is some preliminary evidence from experiments using fused stop-consonant-vowel syllables contrasting in the initial formant transitions only (i.e., place-of-articulation contrasts) that moderate attenuation of the stimulus in one ear has no perceptual consequences (Repp, 1976b). The present study is the first to ask the same question about fused syllables contrasting in VOT.

Although some earlier studies suggest that amplitude relationships are unimportant, it is dangerous to generalize from one situation to another, in view of the different psychoacoustic situations represented by different dichotic experiments. In contrast to syllables that do not fuse, and in contrast to dichotic place contrasts that fuse perfectly, dichotic voicing contrasts are "partially fused," given that they differ only in VOT (Repp, 1978a): The initial aspirated portion of the voiceless stimulus does not fuse with the initial (voiced) portion of the voiced stimulus, but the identical voiced portions in the two stimuli fuse perfectly and create a single auditory image localized between the two ears. Earlier experiments (Repp, 1977a, 1978a) have shown that listeners can perceptually integrate all these stimulus components into a single phonetic percept; yet it is true that a careful listener can easily determine the ear in which the aspiration noise occurs. This--from a psychoacoustic viewpoint--unique situation justifies a separate inquiry into the effects of amplitude relationships on stimulus dominance.

## Method

Subjects. The same subjects as in Experiment I participated.

Stimuli. Each dichotic stimulus consisted of a /da/ presented to one ear and a /ta/ presented simultaneously to the other ear. The stimuli were

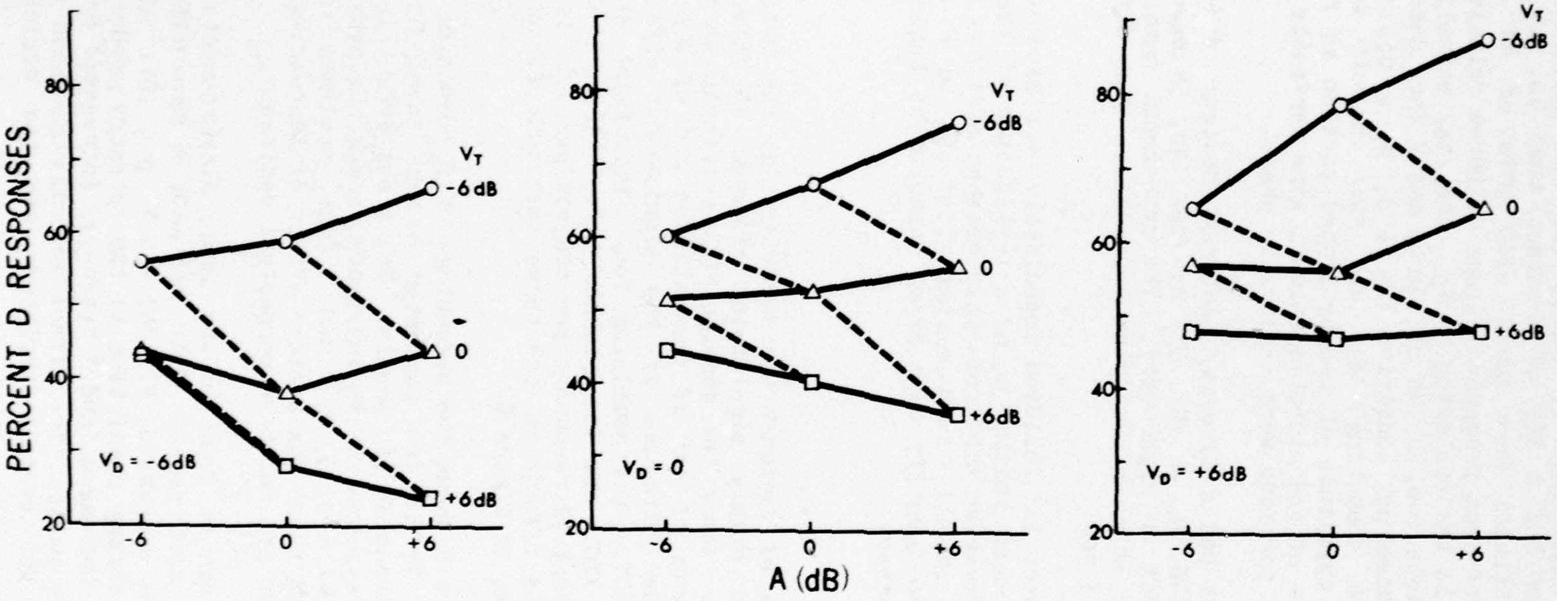


Figure 8: Dichotic voicing contrasts: Percentage of D responses as a function of A, $V_T$, and $V_D$. The dashed lines connect points of equal $A/V_T$ ratio.

identical to those of Experiment I, except for the differences mentioned below. The /da/ always had a VOT of 0 msec; that is, it contained no aspiration. Two /ta/ stimuli were used, with VOTs of 44 and 56 msec, respectively. The /da/ was presented at one of three relative intensities (-6, 0, +6 dB), denoted in the following by $V_D$. The /ta/ stimuli included the same independent variations (-6, 0, +6 dB) in A and V (now denoted $V_T$) as in Experiment I. In the baseline condition ($V_D$ = 0, $V_T$ = 0), the identical voiced portions of the competing /da/ and /ta/ stimuli were equal in amplitude; however, the amplitude of the aspirated portion in /ta/ was about 20 dB lower. The onsets of the dichotic stimuli were perfectly simultaneous, and the identical voiced portions were exactly in phase.

Thus, the experiment had a five-way factorial design: A (-6, 0 +6 dB), $V_T$ (-6, 0, +6 dB), $V_D$ (-6, 0, +6 dB), VOT of /ta/ (44, 56 msec), and ear of presentation (/ta/ in left or right ear). The orthogonal combination of all five factors led to 108 stimuli that were recorded in four different random sequences, with ISIs of 3 sec.

Procedure. Experiment III followed immediately upon Experiment I in each of two sessions, the second session being a replication of the first. Tape recorder channels were reversed electronically between sessions for counterbalancing purposes. All in all, each subject listened to 8 blocks of 108 stimuli. The task was to identify each fused stimulus as beginning with a D or a T, guessing if necessary.

## Results

A five-way analysis of variance was conducted on the response frequencies. There were five highly significant effects; no other effect even approached significance. Among the nonsignificant effects were--quite unexpectedly--the main effects of ear of presentation and of VOT; they will be discussed below, together with one of the significant effects, a triple interaction including VOT. The remaining four significant effects did not include either ears or VOT, and the results were therefore collapsed over these two factors, yielding 32 responses per subject per cell in the remaining three-factor design. The effects of the three amplitude factors, $V_D$, A, and $V_T$, are graphically shown in Figure 8.

Each panel in Figure 8 shows the percentage of D responses as a function of A. In each panel, $V_T$ is the parameter of the three functions (solid lines), whereas $V_D$ increases across panels. The first effect to be seen is a general increase in D responses as $V_D$ increased, comparing data points across the three panels, $F(2,14) = 11.5$, $p < .01$. Thus, contrary to expectations based on earlier studies, there was a clear effect of increasing the amplitude of one stimulus (/da/) on its relative perceptual dominance.

The second effect can be seen in the large displacements of the three solid functions within each panel. This reflects a main effect of $V_T$: D responses increased as $V_T$ decreased, $F(2,14) = 9.7$, $p < .01$. Obviously, this effect contradicts the strong prediction of the category goodness hypothesis: D responses should have decreased (and T responses increased) as $V_T$ decreased, thus increasing the $A/V_T$ ratio. The effect actually obtained indicates that not category goodness but overall amplitude was the decisive factor in dichotic competition.

The third effect displayed in Figure 8 is the fan-shaped pattern of the solid functions in each panel. It represents an A by $V_T$ interaction, $F(4,28) = 4.8$, $p < .01$. This result contrasts with Experiment I, where no such interaction was obtained. When in dichotic competition with a /da/, changes in $V_T$ had a larger effect when A was high than when it was low. Expressed differently, A--the amplitude of the aspirated portion--had different effects depending on the level of $V_T$, the amplitude of the vocalic portion in /ta/. In fact, an increase in A had the expected effect--a reduction in D responses--only when $V_T$ was high (+6 dB). When $V_T$ was low (-6 dB), the effect of A was inverted, higher amplitudes leading to more D (fewer T) responses! Along with the fact that A had no significant main effect, this result refutes even the weaker prediction of the category goodness hypothesis and calls for explanation. (See the Discussion section.)

The interaction just described suggests that a constant $A/V_T$ ratio did not lead to a constant perceptual result. This is confirmed by inspecting Figure 8. Points of constant $A/V_T$ ratio are connected by dashed lines, in analogy with Figure 3. The dashed functions have uniformly negative slopes, indicating that D responses decreased as the overall amplitude of the /ta/ increased, holding $A/V_T$ ratio constant.

The fourth effect shown in Figure 8 is represented by a rotation of the fan-shaped pattern (due to the A by $V_T$ interaction) from a downward orientation in the left-hand panel to an upward orientation in the right-hand panel. This rotation reflects a $V_D$ by A interaction, $F(4,28) = 5.2$, $p < .01$. Thus, the effect of A depended on both $V_D$ and $V_T$, being in the expected direction when $V_D$ was low and $V_T$ high, negligible when $V_D$ equalled $V_T$, and in the opposite direction when $V_D$ was high and $V_T$ was low. This suggests that the subjects' responses were essentially a function of A and $V_D/V_T$ ratio. A comparison of those solid functions representing the same $V_D/V_T$ ratio across different panels in Figure 4 suggests that the direction of the A effect was indeed similar in conditions of equal $V_D/V_T$ ratio, but the absolute response frequencies were not quite the same. There was a small but consistent additional effect of absolute amplitude level: An equal increase in the amplitudes of both competing stimuli (hence, of $V_D$, A, and $V_T$ simultaneously) led to a decrease in D responses.

Before attempting to summarize and interpret this complex pattern of results, the remaining two factors of the experiment, VOT and ear of presentation, require a brief discussion. It came as a surprise that VOT had no effect, since earlier studies had shown that increasing the VOT of the voiceless member of a stimulus pair contrasting in voicing invariably leads to an increase in voiceless responses (Miller, 1977; Repp, 1978a). However, examination of the stimulus specifications revealed that the author had committed an embarrassing error in stimulus synthesis: /ta/ stimuli with a VOT of 56 msec in truth had 12 msec of voicing at onset, superimposed upon the aspiration. These bizarre stimuli, though impossible in articulatory terms, did not sound anomalous and, on the average, turned out to be perceptually equivalent to stimuli with a VOT of 44 msec (which had been properly synthesized). It is interesting to note, however, that individual subjects reacted quite differently to the anomaly: Three subjects (including the author) gave decidedly more T responses when the VOT of /ta/ was nominally 56 msec; thus, they seemed to react primarily to the temporal aspect of VOT. The other five subjects showed just the opposite, and thus seemed to be more sensitive to spectral cues (the presence of periodicity at the onset of the anomalous stimuli).

The statistical analysis showed one significant effect involving VOT, a VOT by $V_D$ by A interaction, F(4,28) = 4.7, $p < .01$. Inspection of the results suggested that the inverted effect of A at the high level of $V_D$ was especially pronounced for stimuli with the anomalous VOT (56 msec). However, the overall pattern of results was similar regardless of VOT, so that this interaction with VOT does not require serious attention.

As to the effect of ear of presentation, there was, as usual, substantial variation from subject to subject. Repp (1977a, 1978a) had reported unusually strong ear dominance effects for fused dichotic voicing contrasts, almost invariably in favor of the right ear. The present results, in contrast to these earlier results, showed no significant overall right-ear advantage, although ear dominance effects were pronounced. The individual results are shown in Table 1.

TABLE 1: Individual ear dominance effects.

| Subject | Sex | Handedness | e' | e |
|---|---|---|---|---|
| BHR | M | R | +0.93 | +0.92*** |
| SB | F | L(familial) | +0.72 | +0.65*** |
| DF | F | R | +0.43 | +0.37*** |
| DK | F | R | +0.29 | +0.27*** |
| JK | M | R | +0.16 | +0.20*** |
| AM | M | L(nonfam.) | -0.11 | -0.13** |
| DW | F | R | -0.50 | -0.45*** |
| MB | F | R | -0.66 | -0.63*** |

**$p < .01$
***$p < .001$

The two ear dominance coefficients reported in Table 1 (e' and e) both range from -1 (total left-ear dominance) to +1 (total right-ear dominance) and generally assume similar values, but only e can easily be tested for significance. The two coefficients and the significance test are described in Repp (1977b). Both coefficients correct for constraints due to stimulus dominance, and e' in particular provides an unbiased estimate of ear dominance.

All eight subjects showed highly significant ear dominance effects. The largest right-ear advantage was shown by the author, in agreement with many previous results; for him, the right ear was almost completely dominant. (His data nevertheless provided information about stimulus dominance: A strongly dominant stimulus in the left ear often overcame the strong right-ear dominance.) The next-largest right-ear advantage was obtained for a familial left-hander--a type of subject who typically does not show large right-ear advantages (Hardyck & Petrinovich, 1977). There were three additional significant right-ear advantages. Of the remaining three subjects, one--a nonfamilial left-hander--showed a small left-ear advantage; the other two subjects were right-handers with large left-ear advantages--a rather puzzling result. Thus, while the present study confirms the proclivity of dichotic voicing contrasts to lead to pronounced individual ear dominance effects, it also increases the urgency of the question whether these effects reflect hemispheric dominance for speech, or perhaps some other kind of lateral asymmetry in auditory processing.

## Discussion

The results of Experiment III clearly refute the category goodness hypothesis of dichotic competition. There was absolutely no evidence in favor of it, suggesting that effects of category goodness were truly absent, not just overcome by more powerful effects of overall amplitude. Although this negative result has been obtained in a situation with very specific psychoacoustic characteristics (partially fused syllables), it certainly raises grave doubts about whether any dichotic stimulus dominance effects can be explained by relative category goodness. Findings previously thought to support this model may ultimately be explainable in psychoacoustic terms.

The key to understanding the complex pattern of results in Experiment III presumably lies in the unique psychoacoustic properties of the dichotic stimuli. Being partially fused, they consisted of a brief unfused part (aspiration noise in one ear, periodic sound in the other) followed by a perfectly fused longer vocalic portion. Changes in the amplitude relationships between the two stimuli (and hence, between the two identical vocalic portions) led to changes in the subjective localization of the fused image. It seems that these localization shifts can explain the perceptual findings.

Consider the case where the amplitude of /da/ is lower than that of the vocalic portion of /ta/ ($V_D < V_T$). In this case, the fused vocalic portion will be localized toward the side where the /ta/ occurred and thus, where the aspiration noise is heard (since it does not fuse with the initial voiced portion of the /da/ in the other ear). In this case, listeners seem to integrate perceptually the aspiration noise with the following vocalic portion and give a majority of /ta/ responses. Moreover, if the amplitude of the aspiration (A) is increased, holding $V_D$ and $V_T$ constant, it has the desired effect of further increasing /ta/ responses. Consider now the case where $V_D > V_T$. Here, the fused vocalic portion will be localized toward the side where the /da/ occurred, away from the aspiration noise. In this case, listeners may find it difficult to integrate the noise with the vocalic portion; therefore, there are few T responses. (This interpretation is reasonable, since informal observations suggest that a noise portion in one ear followed by a vocalic portion in the other ear, i.e., a single /ta/ split between the two ears, is generally not perceived as /ta/ but as a noise in one ear and a vowel in the other.) When A is increased in this situation, it has the paradoxical effect of decreasing T responses. Given that the noise is already segregated from the vocalic portion, an increase in its amplitude probably further increases the perceptual dissociation of the two stimulus components (cf. Dannenbring & Bregman, 1976). The case of $V_D = V_T$ falls between these two extremes. Here, the fused vocalic portion is localized in the midline, and the opposed effects of changes in A on the voicing decision and on the perceptual dissociation of noise and vocalic portions seem to cancel out, since A has little effect on the responses.

Thus, the results can be explained by taking account of the relative positions of the stimulus portions in subjective space. The four principal effects in the data described earlier ($V_D$, $V_T$, $V_D$ x A, $V_T$ x A) are really only two: There is an effect of $V_D/V_T$ ratio (that is, of closeness in auditory space of the aspirated portion and the fused vocalic portion) and an interaction between $V_D/V_T$ ratio and A (that is, the effect of A depends on whether the aspiration can be integrated with the fused vocalic portion). The small additional effect of overall stimulus amplitude (more T responses when amplitude was uniformly increased) probably reflects a certain amount of interference of the initial periodic portion of /da/ with the perception of the simultaneous aspiration noise in the other ear, and this interference decreased at higher stimulus intensities. The present data do not contradict earlier dichotic studies that found no amplitude effects, once it is realized how unique the psychoacoustic properties of partially fused voicing contrasts are.

The results of Experiment III provide an interesting example of how spatial separation can lead to perceptual dissociation. Related findings in the literature include the perceptual difficulties encountered in the perception of continous speech alternated between the two ears (Cherry & Taylor, 1954; Huggins, 1964), in numerosity judgments of dichotically alternated pulse trains (Axelrod, Guzy, & Diamond, 1968; Huggins, 1974), and in dichotic periodicity perception (Huggins, 1976). Another related paradigm in which perceptual integration across ears seems to be more successful than in the present experiment is the so-called spectral/temporal fusion (Cutting, 1976) observed when the second-formant transition of a stop-consonant-vowel syllable is presented to one ear and the remainder of the stimulus to the other. Rand (1974) has shown that stimulus intelligibility can even be improved by dichotic separation of these stimulus components. The factors that determine whether integration of spatially and temporally disparate inputs does or does not occur probably include spectral disparity and amplitude in addition to the amounts of spatial and temporal separation. This related research is cited in order to point to the general class of psychoacoustic phenomena that subsumes the peculiar situation created by dichotic voicing contrasts. The present results demonstrate an often neglected problem: The dichotic presentation of two speech stimuli may lead to complex and possibly unique binaural interactions that are likely to constitute major determinants of ear and stimulus dominance effects.

## REFERENCES

Axelrod, S., Guzy, L. T., & Diamond, I. T. Perceived rate of monotic and dichotically alternating clicks. *Journal of the Acoustical Society of America*, 1968, *43*, 51-55.

Bailey, P. J., & Summerfield, Q. Some observations on the perception of [s]+stop clusters. *Haskins Laboratories Status Report on Speech*

Research, 1978, SR-53 (vol. 2), 25-60.

Cherry, E. C., & Taylor, W. K. Some further experiments upon the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America, 1954, 26, 554-559.

Cullen, J. K., Jr., Thompson, C. L., Hughes, L. F., Berlin, C. I., & Samson, D. S. The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain and Language, 1974, 1, 307-322.

Cutting, J. E. Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review, 1976, 83, 114-140.

Dannenbring, G. L., & Bregman, A. S. Stream segregation and the illusion of overlap. Journal of Experimental Psychology (Human Perception and Performance), 1976, 2, 544-555.

Divenyi, P. L., & Danner, W. F. Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. Perception and Psychophysics, 1977, 21, 125-142.

Hardyck, C., & Petrinovich, L. F. Left-handedness. Psychological Bulletin, 1977, 84, 385-404.

Homick, J. L., Elfner, L. F., & Bothe, G. G. Auditory temporal masking and the perception of order. Journal of the Acoustical Society of America, 1969, 45, 712-718.

Huggins, A. W. F. Distortion of the temporal pattern of speech: Interruption and alternation. Journal of the Acoustical Society of America, 1964, 36, 1055-1064.

Huggins, A. W. F. On perceptual integration of dichotically alternated pulse trains. Journal of the Acoustical Society of America, 1974, 56, 939-943.

Huggins, A. W. F. Is periodicity detection central? Journal of the Acoustical Society of America, 1976, 59 (Suppl. No. 1), S52(A).

Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.

Lisker, L. Is it VOT or a first-formant transition detector? Journal of the Acoustical Society of America, 1975, 57, 1547-1551.

Lisker, L., & Abramson, A. S. A cross-language study of voicing in initial stops: acoustical measurements. Word, 1964, 20, 384-422.

Lisker, L., & Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague 1967, Pp. 563-567. Prague: Academia, 1967.

Lisker, L., & Abramson, A. S. Distinctive features and laryngeal control. Language, 1971, 47, 767-785.

Massaro, D. W., & Cohen, M. M. The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America, 1976, 60, 704-717.

Massaro, D. W., & Cohen, M. M. Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. Perception and Psychophysics, 1977, 22, 373-382.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times. Journal of the Acoustical Society of America, 1976, 60, 410-417.

Miller, J. L. Properties of feature detectors for VOT: The voiceless channel of analysis. Journal of the Acoustical Society of America, 1977, 62, 641-648.

Pisoni, D. B. Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Pompino, B., Rilhac-Sutter, M., Simon, A., & Sommer, R. Auditorische Faktoren der Gewichtung bei psychoakustischer Fusion. Forschungsberichte des Instituts fuer Phonetik und sprachliche Kommunikation, 1977, 8, 97-120. (University of Munich.)

Rand, T. C. Dichotic release from masking for speech. Journal of the Acoustical Society of America, 1974, 55, 678-680.

Repp, B. H. Identification of dichotic fusions. Journal of the Acoustical Society of America, 1976a, 60, 456-469.

Repp, B. H. Discrimination of dichotic fusions. Haskins Laboratories Status Report on Speech Research, 1976b, SR-45/46, 123-139.

Repp, B. H. Dichotic competition of speech sounds: The role of acoustic stimulus structure. Journal of Experimental Psychology: Human Perception and Performance, 1977a, 3, 37-50.

Repp, B. H. Measuring laterality effects in dichotic listening. Journal of the Acoustical Society of America, 1977b, 62, 720-737.

Repp, B. H. Stimulus dominance and ear dominance in the perception of dichotic voicing contrasts. Brain and Language, 1978a, 5, 310-330.

Repp, B. H. Stimulus dominance in fused dichotic syllables. Haskins Laboratories Status Report on Speech Research, 1978b, SR-55, 133-148.

Repp, B. H. Categorical perception of fused dichotic syllables. Haskins Laboratories Status Report on Speech Research, 1978c, SR-55/56, 149-161.

Repp, B. H. Stimulus dominance and ear dominance in fused dichotic speech and nonspeech stimuli. Haskins Laboratories Status Report on Speech Research, 1978d, SR-55/56, 163-179.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of temporal cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

Speaks, C., & Bissonette, L. J. Interaural-intensive differences and dichotic listening. Journal of the Acoustical Society of America, 1975, 58, 893-898.

Summerfield, A. Q., & Haggard, M. P. Perceptual processing of multiple cues and contexts: effects of following vowel upon stop consonant voicing. Journal of Phonetics, 1974, 2, 279-295.

Summerfield, Q., & Haggard, M. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 1977, 62, 435-448.

Winitz, H., LaRiviere, C., & Herriman, E. Variations in VOT for English initial stops. Journal of Phonetics, 1975, 3, 41-52.

Zue, V. W. Acoustic characteristics of stop consonants: A controlled study. Lincoln Laboratories Technical Report 523. M.I.T., Lexington, Massachusetts, 1976.

## FOOTNOTES

[1] It is necessary to distinguish between two meanings of the term VOT: In articulation, it denotes the articulatory gesture of laryngeal adjustment which underlies the manifold acoustic cues. Following common usage, however,

VOT at the acoustic level refers only to the temporal separation between release and voicing onset; it does not include spectral voicing cues, such as $F_1$ onset, that can be independently manipulated in synthetic speech.

[2] Dissertation research currently being conducted by Barbara Moslin at Brown University may be the only exception. At this time, I have not seen a sufficiently detailed account of that work to be able to compare her findings with my data.

[3] Arthur Abramson and Leigh Lisker, personal communication.

[4] Pastore, R. E. Psychoacoustic factors in speech perception. To appear in a book edited by P. D. Eimas and J. L. Miller. This chapter provides an excellent review of the relevant issues.

[5] Virginia Mann was instrumental in guiding me toward this interpretation.

[6] It is entirely possible that burst intensity constitutes an important cue when aspiration is absent, as in the voiced-voiceless distinction for unaspirated stops in certain languages that feature such a distinction.

# STOP CONSONANT PLACE PERCEPTION WITH SINGLE-FORMANT STIMULI: EVIDENCE FOR THE ROLE OF THE FRONT-CAVITY RESONANCE*

G. M. Kuhn+

Abstract: The third formant and the second formant were found on average to cue the place of articulation of intervocalic stop consonants equally well when the stop consonants occurred before the vowel /i/. This result and others provide some support for the notion that the fundamental resonance of the front cavity plays an important role in the perception of the phonetic dimension of place of articulation.

## I. INTRODUCTION

In an earlier report (Kuhn, 1975), we presented spectrographic evidence for articulatory rationalizations of some of the acoustic cues for speech, to emphasize the possible role of the fundamental resonance of the front cavity, "the front cavity resonance," in the perception of the phonetic dimension of place of articulation.[1] It appeared that familiar contributions of $F_2$, $F_3$ and stop bursts to place perception with synthetic speech could be reinterpreted to reflect contributions of the front cavity resonance. Also, several anomalous results with the same formants and bursts in synthetic speech appeared to be interpretable in terms of the front cavity resonance.

Two problems arose from that report. A first problem was that, since formant-cavity affiliations can change, it was important to show that different formants could contribute highly to the perception of place, depending on their degree of affiliation with the front cavity. Several results from speech synthesis were cited. Many well-known results (e.g., Delattre, Liberman, Cooper, & Gerstman, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954) indicated that $F_2$ contributes highly to the perception of stop consonant place of articulation in stop-vowel syllables where the front cavity resonance is close to $F_2$. However, only one result (from Harris, Hoffman, Liberman, Delattre, & Cooper, 1958), was interpreted to show that $F_3$ contributes highly to the perception of place of articulation in stop-vowel

*This paper will be published in the Journal of the Acoustical Society of America.

+Institute for Defense Analyses, Princeton, New Jersey.

Acknowledgment: This is a revised version of G. M. Kuhn's Ph.D. thesis (1978) for the University of Connecticut, Department of Linguistics. The research reported was performed at Haskins Laboratories under NIH Grant HD01994 and Veterans Administration Contract V101(134)P-342.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

syllables where the front cavity resonance is close to $F_3$. Furthermore, some stimuli constructed according to the formant-frequency paradigm of that experiment may not have been representative of natural speech. Therefore, it seemed a good idea to design an experiment whose results could show a change in the ability of the natural $F_2$ and $F_3$ to provide stop consonant place perception that is in the direction of the expected cavity affiliation changes.

A second problem arising from that earlier report had to do with our claim that a special kind of speech, called "fricative speech," is highly intelligible. Fricative speech is produced by moving the lips, tongue and jaw while maintaining enough tongue constriction to cause a turbulent source to be created at the palate. If constriction is imposed elsewhere, as with labial or apical closure, then the sound source can move to another point in the vocal tract, as it does in normal speech. Unlike normal speech, however, the sound source is always located at the point of major constriction. This means that fricative speech vowels are produced with a much more constricted vocal tract than normal speech vowels.

Our interest in the claim about the intelligibility of fricative speech arose from the fact that fricative speech is essentially single formant, front cavity resonance speech. Its single formant, $F_f$ ("F front"), is close to the frequency of a prominent formant or group of formants for the corresponding articulations in normal speech. See, for example, the comparison of fricative and normal speech stimuli in Figure 1. The problem with the claim was that there were no experimental results to support it.

In other words, whether we were talking about the predicted manifestations of the front cavity resonance in normal or fricative speech, more data were needed to support the perceptual claims of the front cavity theory. We decided to assess the accuracy of place perception for single-formant stimuli that copy either $F_2$ or $F_3$ from normal speech, or the front cavity resonance, $F_f$, from fricative speech.

The corpus consisted of the six stop consonants of American English embedded in 16 intervocalic environments formed by the inclusive combination of /iauɝ/, yielding a total of 96 vowel + stop consonant + vowel (VCV) utterances. The intervocalic environment was chosen because it gives both closure and release information about the consonant (Ohman, 1966), and because it gives the subjects a chance to anticipate the consonant. The vowels were chosen in part because they include extreme positions of the first three formant frequencies of the vowels of American English. Furthermore, in normal voiced speech, the vowel /i/ is a close front vowel, for which the $F_3$ region can be more closely associated with the front cavity than $F_2$, whereas the vowels /a/, /u/ and /ɝ/ are back vowels, for which, in a highly constricted articulation, the $F_2$ region is predicted to be more closely associated with the front cavity than $F_3$.

Two important qualifications need to be made about the description of the vowels in the last paragraph. First, the use of the terms "front" or "back" is intended to imply only that the position of the major tongue constriction is anterior or posterior to the position at which the $F_2$ and $F_3$ formant-cavity

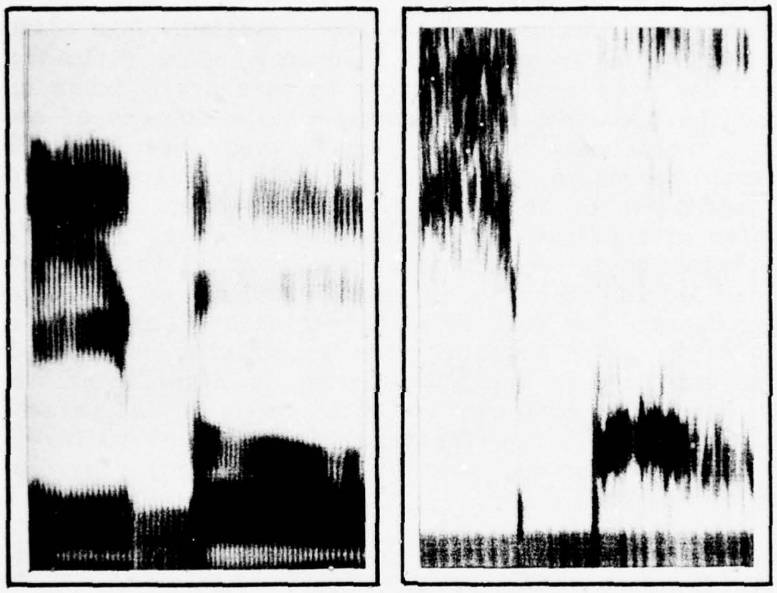


Figure 1: Spectrograms of the utterance /ibu/ spoken in normal speech (left) and fricative speech (right). Note the correspondences in spectral energy across the two tokens.

affiliations change for a highly constricted vowel configuration. See, for example, the discussion of cavity affiliations in Stevens and House (1956), or the nomograms of Fant (1960). Second, the degree to which the formants enjoy the indicated association with the front cavity will vary of course with the amount of constriction, and, in the case of /i/, can vary with the exact placement of the constriction. Thus, the formant-cavity affiliations may be stronger shortly after a stop release than during more open parts of a vowel articulation, and, it is possible for the front cavity to be more closely associated with $F_4$ of /i/ if the tongue constriction is both extreme and extremely advanced.

The stop consonants provide the kind of extreme constriction that can produce a well-defined front cavity and front cavity resonance. For velar stop consonants, the release tends to have a strong component near the fundamental resonance frequency of the front cavity. This component co-varies in frequency with the front-cavity resonance of a following vowel (Fant, 1960). For apical stop consonants, the release again tends to have a strong component near the frequency of the fundamental resonance of the front cavity, this time at a relatively high frequency, i.e., near the frequency of the third or fourth resonance mode of the whole vocal tract (Stevens, 1972, pp. 51-66), regardless of the following vowel. Then, if the tongue tip drops and the location of the sound source changes abruptly, a lower resonance mode of the whole vocal tract may become strongly affiliated with the front cavity resonance. For apical stops before back vowels like /u/, for example, the front cavity resonance can fall from the region of $F_3$ or $F_4$ to a low frequency in the region of $F_2$. For bilabial stop consonants, no front-cavity quarter-wave resonance can appear until the mouth is significantly open. Then, a front cavity resonance component of the release will rise in frequency, usually from a relatively low frequency in the region of $F_2$. For bilabial stops before a front vowel such as /i/, this frequency rise can continue even to the region of $F_3$ (Kuhn, 1975).

A third problem was created by our decision to compare the front cavity resonance with formants indicated by number. This problem had to do with the treatment of the burst. The burst can be defined as the rapid onset of energy that tends to occur during the first few milliseconds after release of a stop consonant. To relate the burst to a following transition, one could treat it as the initiation of an acoustic feature of rapid spectrum change at consonant release (Stevens, 1973). Another way to relate the burst to a following spectral transition would be to think of it as the initial manifestation of the front cavity resonance, at least for the apical and velar stops, where there can be a well-defined cavity in front of the constriction.

Given this latter notion of the burst, the problem in comparing $F_f$ with $F_2$ and $F_3$ becomes apparent. Front cavity resonances prevail in the burst of stop consonants (Fant, 1960, 1969; Heinz & Stevens, 1961: Stevens, 1968); and after the burst, our fundamental front cavity resonance may change frequency very discontinuously, exciting a different resonance mode of the whole vocal tract if the location of the sound source changes suddenly, e.g., in the transition to a back vowel after an apical stop. However, the classic usage of the terms "$F_2$" and "$F_3$" restricts them to moments in time when they can be traced continuously toward characteristic vowel formant frequencies. This is particularly true for the synthetic $F_2$ and $F_3$ on which many of the claims about cues to place of articulation are based. If the front cavity resonance of fricative speech is evident in the release burst, and if we strip off the burst in our tests, then we may get a false picture of the ability of the front cavity resonance to provide place perception. On the other hand, we do not want to report that $F_f$ did better than $F_2$ or $F_3$ simply because the latter were defined not to include the analogous burst energy. For this reason, we decided to assess the accuracy of stop consonant place perception for $F_f$, $F_2$ and $F_3$ both with and without any detected stop consonant release burst.

Our predictions about the outcome of the experiment were guided by the expected formant-cavity affiliations of the vowel, and by the hypothesis that the importance of a formant in providing place perception varies directly with its front cavity affiliation. The predictions were, that before the front vowel /i/, $F_f$ would provide place identification scores equal to those for $F_3$, since both were expected to have significant front cavity affiliation, and both would do better than $F_2$, which was expected to have much less affiliation with the front cavity. On the other hand, before the back vowels /a/, /u/, and /ɝ/, $F_f$ would provide, we predicted, place identification scores equal to those for $F_2$, and both would do better than $F_3$, which was expected to have much less affiliaton with the front cavity.

These predictions emphasize the role of the following vowel context. However, effects of a preceding vowel on the acoustic structure of an intervocalic stop consonant have been shown (Ohman, 1966). Therefore, we looked for an effect of the preceding vowel context as well (see below).

It is clear that we have made the predictions depend only on the vowel, and not, for example, on the place of articulation, although we expect the front cavity resonance to appear at or after the burst, and even to change its affiliation from one formant to another during some stop-vowel transitions (e.g., apical stop to /u/, or bilabial stop to /i/). The reason for doing this is simply to highlight the effect of front-cavity affiliation on the overall or average contribution of these formants to place perception. We do prefer to emphasize the global manifestation of the front cavity resonance in a stop release, independent of the manner of excitation: from the burst, through any frication, any aspiration and into any voiced formant transition. More will be said about the hypothesized cue value of the global manifestation of the front cavity resonance, and about its relation to some other hypotheses on place perception, below.

## II. METHODS

### The Experimental Design

Two experiments were carried out, a preliminary experiment with natural stimuli, and a main experiment with synthetic stimuli copied from the natural. The preliminary experiment with natural stimuli consisted of 2 conditions, one with normal speech, and the other with fricative speech. The main experiment with synthetic stimuli consisted of 8 conditions. All of the experimental conditions are listed in Table 1. Note the use of "F" followed by a subscript

THE PRELIMINARY EXPERIMENT: NATURAL STIMULI

| | Condition |
|---|---|
| 1. | Normal speech |
| 2. | Fricative speech |

THE MAIN EXPERIMENT: SYNTHETIC STIMULI

| | Condition | Excitation | Description |
|---|---|---|---|
| i. | P | varied | practice |
| 1. | F123 | normal | (burst +) $F_1 + F_2 + F_3$ |
| 2. | F23 | aperiodic | (burst +) $F_2 + F_3$ |
| 3. | F2 | aperiodic | (burst +) $F_2$ |
| 4. | F2- | aperiodic | $F_2$, no bursts |
| 5. | F3 | aperiodic | (burst +) $F_3$ |
| 6. | F3- | aperiodic | $F_3$, no bursts |
| 7. | FF | aperiodic | (burst +) $F_f$ |
| 8. | FF- | aperiodic | $F_f$, no bursts |

Table 1: A list of the conditions for the preliminary experiment with natural stimuli, and for the main experiment with synthetic stimuli.

(e.g., "$F_2$") to denote a formant to be tested in an experimental condition, while "F" followed by one or more characters on the line (e.g., "F2") denotes the experimental condition itself. Also, it should be clear that we have been using the term "normal" speech in contrast with fricative speech, while "natural" stimuli are contrasted with synthetic stimuli. Aperiodic excitation was used for all of the single-formant stimuli because fricative speech uses aperiodic excitation. The notation "(burst +)" means "burst if any, plus." More is said about "normal" excitation below.

## Stimulus Preparation

The first step in the process of stimulus preparation was to create two analog tape recordings of the 96 natural VCV̲'s, the one spoken in fricative speech and the other in normal speech. A randomization of the 96 VCV's was read aloud in fricative speech, in an anechoic chamber, into a condenser microphone, and recorded on a high-quality tape recorder located outside the booth. A recording that included the 96 normal speech VCV̲'s, made by the same speaker under identical circumstances, was available from an earlier experiment (Kuhn & McGuire, 1974). In both recordings, heavy or emphatic stress was placed on the second syllable.

The next step in stimulus preparation was to obtain digital waveforms and spectra from the analog tape recordings. The recording of the fricative speech VCV̲'s was played back from a tape recorder along two paths to a multiplexer and A/D converter, and finally to a computer (Honeywell DDP-224). One path took the signal through a pre-emphasis circuit and a 90 Hz - 4 kHz band-pass filter. This waveform path delivered 10,000 12-bit waveform samples per second to the computer. The other path took the signal through a different pre-emphasis circuit and a 5 kHz low-pass filter, and then into a real-time spectrum analyzer (Federal Scientific UA6A). This spectrum path delivered 10,000 8-bit spectrum amplitude samples per second to the computer, representing the speech band to 4920 Hz in adjacent 40 Hz channels, one frequency scan following another every 12.8 msec.[2] With the multiplexer switching between the waveform and spectrum paths 20,000 times per second, the system delivered the natural fricative speech waveform and spectrum data "simultaneously" and in real time, to the computer. These data were transferred in real time to digital magnetic tape, and ultimately to disc. Disc files of data for the 96 natural normal speech VCV̲'s were available from the same experiment mentioned above (Kuhn & McGuire, 1974).

Once the waveform and spectrum data for both the 96 natural fricative speech stimuli and the 96 natural normal speech stimuli were available on-line, the next step was to estimate the frequencies and amplitudes of the bursts and formants for the synthetic stimuli. The first sets of synthetic stimuli to be created were the single-formant fricative speech stimuli, and the three-formant normal speech stimuli.

For the fricative speech stimuli, each amplitude spectrum was filtered (triangular window, 320 Hz) and compressed (by one-half of that spectrum's average channel amplitude). A dot-density spectrogram and peak estimates for the stimulus were then displayed on a storage scope (see Figure 2). Where many peaks appeared close together, a display of the local peak of maximum

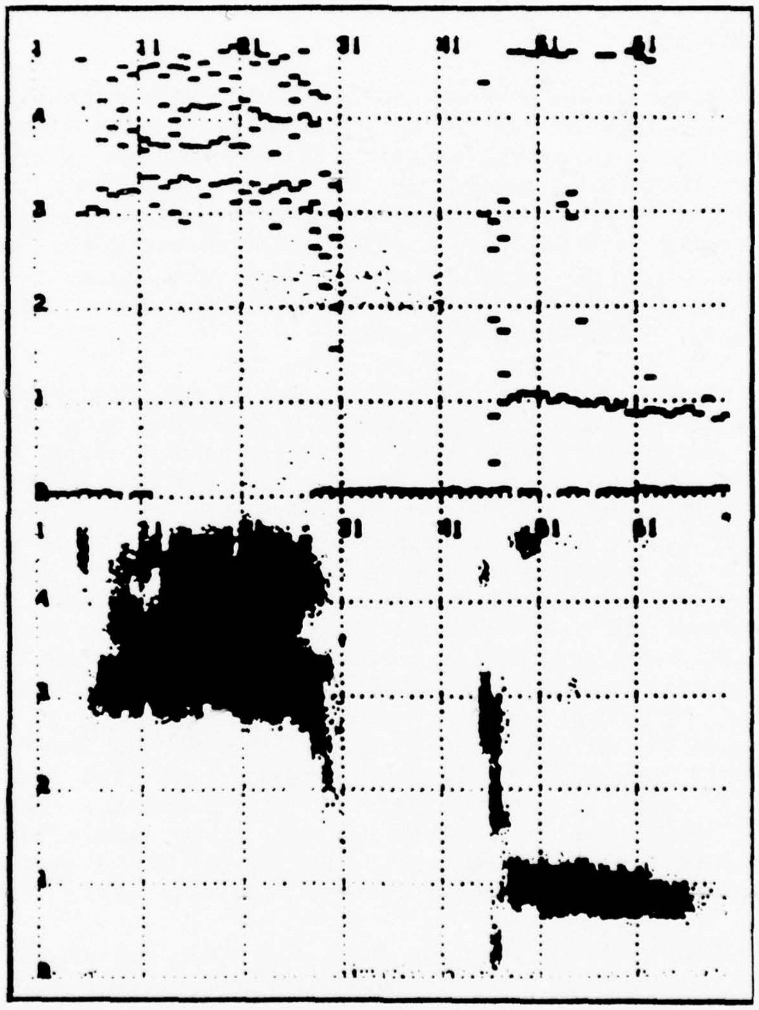


Figure 2: Dot-density spectrogram (bottom) and peak estimates (top) for the natural fricative speech /ibu/.

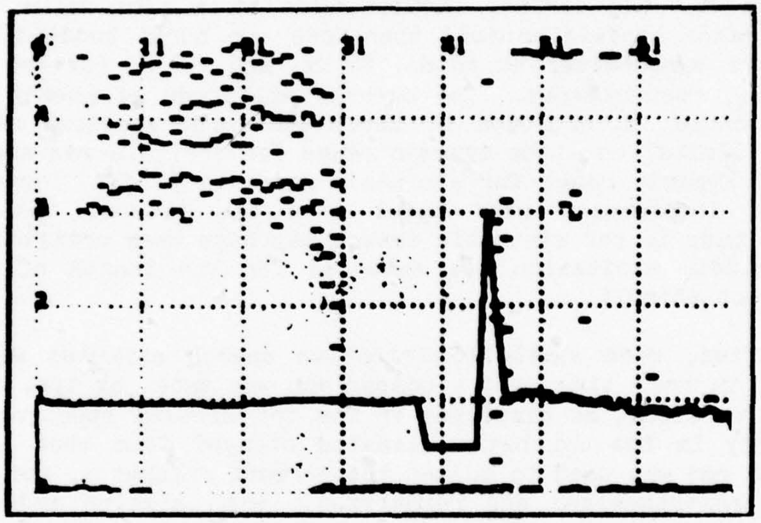


Figure 3: The $F_2$ parameter of the synthesizer was used to make part of $F_f$ in the single-formant synthetic fricative speech /ibu/. For time frames 1-38, $F_2$ was traced by hand at approximately 1 kHz. For the time frames 39-70, $F_2$ was determined with computer assistance, as follows. The operator traced a line below the spectral peak to be used as $F_2$. For time frames 39-44, there was no higher spectral peak so the computer substituted the minimum frequency of the $F_2$ parameter. For time frames 45-70, a higher spectral peak was found and its frequency was kept as the frequency of $F_2$. Note that the amplitude of $F_2$ would not be turned on until frame 45: the $F_3$ parameter of the synthesizer was used to simulate $F_f$ in this stimulus up to the stop closure.

amplitude was used to help in peak selection. A graph pen was then used in conjunction with the storage scope to underline the sequence of spectral peaks to be used for the single-formant synthetic stimulus3 (see Figure 3). Note that this means that a single formant was used to simulate the burst, if any, as well. The indicated peak frequencies were then automatically converted to control parameter values for the Haskins Laboratories parallel resonance synthesizer (Epstein, 1965). The peak frequencies were converted to the nearest $F_3$ code (step size 170 Hz) in the syllables with /i/. The peak frequencies were also converted to the nearest $F_3$ code in other syllables if the $F_2$ range was exceeded, e.g., if the burst frequency was above the $F_2$ range. Elsewhere the peak frequencies were converted to the nearest $F_2$ code (step size 79 Hz). The formant amplitude parameters were set at a constant non-zero value for the desired formant, and at a zero value for all other formants. Formant bandwidths (and therefore our burst bandwidth) were fixed by the hardware synthesizer at 60 Hz, 80 Hz, and 100 Hz for the synthesizer's $F_1$, $F_2$, and $F_3$, respectively. The overall amplitude parameter (step size 3 dB) provided output attenuation to match the input spectrum amplitude, with the following limitation: the dynamic range for analysis was approximately 48 dB, while the dynamic range for synthesis was only 30 dB. Therefore, if very low amplitude information was copied from the natural speech, then its relative amplitude in the synthetic speech may have been artificially boosted. Finally, aperiodic excitation was selected for the length of the synthetic fricative speech stimuli.

So specified, each synthetic fricative speech stimulus was synthesized and analyzed, in real time, and a comparison was made, by eye, of the natural and synthetic versions, as displayed in the dot density spectrograms. If the spectral energy in the synthetic stimulus strayed from that in the natural one, the graph pen was used to adjust the formant frequency, and the synthetic stimulus was resynthesized and reanalyzed. With similar adjustment of the overall amplitude parameter, close (burst +) single-formant approximations to the natural fricative speech stimuli were obtained in just a few iterations. Hand smoothing of the formant frequencies and of the amplitude parameters was used to eliminate the "watery sound" due to the lack of continuity in the values obtained by automatic peak picking. See the sample synthetic fricative speech stimulus in Figure 4.

The procedure for developing the three-formant synthetic normal speech stimuli was similar to that for developing the single-formant synthetic fricative speech stimuli. Dot density spectrograms of the filtered and compressed spectral data were displayed on the storage scope, along with spectral peak estimates, and the graph pen was used to select the three formant frequencies. A single formant was used to represent the burst, if any. The synthesizer's overall amplitude parameter provided output attenuation for time frames excited with an aperiodic source (i.e., those where there was no visible $F_1$), and for time frames excited with a periodic source (i.e., those where there was a visible $F_1$). All amplitude parameters were turned off during the stop gap, for both voiceless and voiced stop consonants, so that place perception could not be assisted by the resonance pattern during closure. Then, the stimulus parameter values were output to the synthesizer, the sound from the synthesizer was sent to the spectrum analyzer, and the spectral analyses for the natural and synthetic versions

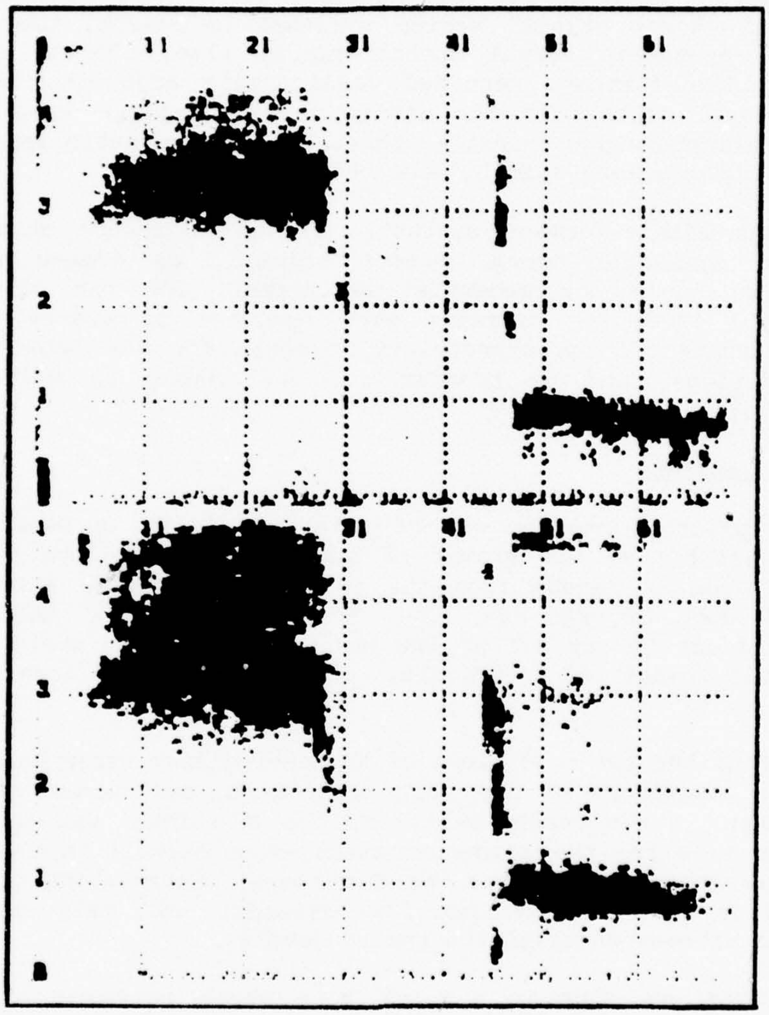


Figure 4: Spectrograms of the synthetic (top) and natural (bottom) versions of the fricative speech /ibu/.

were compared. The graph pen was used to touch up and to smooth the parameters.

Differences in the procedure for creating the three-formant synthetic normal speech stimuli arose for two reasons. First, a fundamental frequency contour had to be imposed on all 96 stimuli. The imposed $F_0$ (fundamental frequency) contour was flat at 120 Hz for the first 548 msec of each stimulus, and fell at a rate of 98 Hz/sec for the remainder of each stimulus. Since the natural normal speech stimuli varied somewhat in length, the final audible pitch of the synthetic normal speech varied also. Second, the relative amplitude of the formants required considerable adjustment to achieve an acceptable visual match with the natural levels.[4] After several synthesis-analysis iterations, close (burst +) three-formant synthetic approximations to the natural normal speech stimuli were obtained.

Given the single-formant synthetic fricative speech stimuli, and the three-formant synthetic normal speech stimuli, as traced with computer assistance, we could now generate the stimuli for the other conditions automatically. Files of commands were executed to create the remaining stimuli by turning off the appropriate formants for the double- and single-formant conditions, and by turning off the bursts in addition, for the burstless conditions.

## Experimental Sessions

For the practice session of the main experiment, a tape recording was made that consisted of six groups of ten stimuli. The ten stimuli in each group were chosen at random from the following synthetic stimulus sets, in order: F123, F23, F2, F3, F2-, F3-. These stimuli were chosen so that the subjects would not be put off at the start, and so they would have practice with the full range of difficulty provided by the conditions of the experiment.

For each of the two conditions of the preliminary experiment and for each of the eight conditions of the main experiment, one format was used in the tape recordings. A new randomization of the 96 stimuli was created with two presentations in a row per stimulus, twenty-four stimuli in a block, and four blocks. Inter-presentation time was one second, inter-stimulus time was two seconds, and inter-block time was five seconds. All main experiment tapes were recorded without changing the record levels.

All testing was done in a quiet room where subjects listened to the stimuli over headphones. Comfortable listening levels were used in the two conditions of the preliminary experiment. Identical comfortable listening levels were used in Conditions P and F123 of the main experiment, and in the final seven conditions of the main experiment (the seven with aperiodic excitation).

In the preliminary experiment the first condition to be run was the natural normal speech condition. The four volunteer subjects were asked to write the symbol for the stop consonant they thought they heard, taking their responses from the set of symbols /bdgptk/ and guessing if necessary. The second condition to be run was the natural fricative speech condition. Although the main interest in this second condition was still to assess the accuracy of stop consonant place perception, the possibility of obtaining data on fricative speech vowel identification proved too much of a temptation. Subjects were instructed to write the initial vowel, the stop consonant, and the final vowel that they thought they heard, taking their responses from the set of symbols /bdgptkiauɝ/ and guessing if necessary.

In the main experiment, all 10 subjects listened to Conditions P, F123, and F23 at the beginning of the experiment. Then, half of the subjects listened to the six single-formant conditions in the order FF, F2, F3, F3-, F2-, FF-, while the other half of the subjects listened to the single-formant conditions in the order F3-, F2-, FF-, FF, F2, F3. This experimental design did not provide complete counterbalancing for ordering effects, but it made sure that the mean position for conditions with the $F_2$, $F_3$ or $F_f$ was the same across all subjects.

At the beginning of the main experiment, the subjects were read the following text:

> You will be listening to vowel + stop consonant + vowel stimuli. Examples are /ɝtu, iba, agi/ (in normal speech), or /ɝtu, iba, agi/ (in fricative speech). They are synthetic stimuli that copy different parts of the information found in natural speech. The purpose of these experiments is to find out what parts of natural speech sounds contribute most to the perception of stop consonants. Your task in each condition of this experiment will be to listen to two repetitions of each stimulus and then to write the letter b, d, g, p, t or k, for the stop consonant that you think you heard. If you are unsure what consonant you heard, then make a guess. Please, leave no answers blank.

Then the subjects were asked if they had any questions. After any questions, the first four experimental conditions (practice session included) were run, followed by a five minute break. Then the next three experimental conditions were run, followed by another five minute break. Then the final two conditions were run, and the subjects were paid and dismissed.

## III. <u>EXPERIMENTAL RESULTS</u>

### The Preliminary Exeriment

In the total of 384 responses to the natural normal speech condition, there were only two errors, made back-to-back by one subject who lost his place. On the basis of this result it was assumed that the natural normal speech stimuli were highly identifiable and that any decrement in identifiability of the synthetic versions would be attributable to the restrictions imposed by synthesis.

The responses to the natural fricative speech condition are presented in Table 2 as a confusion matrix summed over all subjects and both vowel

NATURAL FRICATIVE SPEECH

RECEIVED

|  |  | b | p | d | t | g | k |
|---|---|---|---|---|---|---|---|
| SENT | b | 21 | 43 |  |  |  |  |
|  | p | 3 | 61 |  |  |  |  |
|  | d |  | 4 | 20 | 37 | 1 | 2 |
|  | t |  | 3 | 2 | 55 |  | 4 |
|  | g |  |  |  |  | 18 | 46 |
|  | k |  |  |  |  | 4 | 60 |

RECEIVED

|  |  | i | a | u | ɝ |
|---|---|---|---|---|---|
| SENT | i | 192 |  |  |  |
|  | a |  | 61 | 3 | 128 |
|  | u |  |  | 192 |  |
|  | ɝ |  | 44 | 15 | 133 |

**Table 2: The consonant (top) and vowel (bottom) confusion matrices for the natural fricative speech condition of the preliminary experiment.**

positions. The main result is that place identification for the natural fricative speech stop consonants averaged 96%. In addition, vowel identification averaged 100% for /i/ and /u/. Vowel identification averaged only 32% for /a/ and 69% for /ɝ/, as the subjects did not distinguish reliably between these two intended vowels. Finally, voicing identification for the stop consonants averaged 63%.

## The Main Experiment

Percent place identification scores were computed for the responses to the main experiment, and in addition, the nonparametric, Wilcoxon matched-pairs signed ranks test (Wilcoxon, 1945; Siegel, 1956) was used to test for significant differences across sets of scores. The Wilcoxon was used to compute the significance of *within-condition* differences between all places of articulation and between all vowel contexts. Although all place and vowel differences were analyzed, it was not meaningful to compare place scores across all eight by eight conditions. Table 3 has an "x" at every position of the *across-condition* matrix where the Wilcoxon test was applied. Whenever a probability is mentioned in the text it will refer, unless otherwise stated, to the one-tailed probability of a Wilcoxon z.

Table 3: A condition-by-condition matrix for the main experiment. An "x" appears in the matrix wherever a Wilcoxon comparison was made, of the place identification scores from the indicated conditions.

ACROSS-CONDITION COMPARISONS

| | F23 | F2 | F3 | FF | F2- | F3- | FF- |
|---|---|---|---|---|---|---|---|
| F123 | x | x | x | x | | | |
| F23 | | x | x | x | | | |
| F2 | | | x | x | x | | |
| F3 | | | | x | | x | |
| FF | | | | | | | x |
| F2- | | | | | | x | x |
| F3- | | | | | | | x |

Figures 5, 6, and 7 present the place identification scores from three different analyses of the responses. Figure 5 presents the place identification scores from the analysis of all stimuli by following vowel. Figure 6 presents the place identification scores from the analysis of only those stimuli that were synthesized with a burst, again by following vowel. Finally, Figure 7 presents the place identification scores from the analysis of all stimuli by preceding vowel. In each of the three figures, from left to right at the top, are the graphs for Conditions F123, F2, F3 and FF. From left to right at the bottom, are the graphs for Conditions F23, F2-, F3- and FF-. From left to right within each graph, the place identification scores are shown for each vowel context: /i/, /a/, /u/, and /ɝ/. In Figures 5 and 6, these are the following vowels, since the place scores were averaged over the preceding vowels; in Figure 7 these are the preceding vowels, since the place scores were averaged over the following vowels. Next to the name for each condition, is the overall percent place identification score for the condition, i.e., percent place identification averaged over all three intended places of articulation and all 16 vowel contexts. As the legend indicates, a solid line shows the scores for the bilabial category, i.e., percent /b/ or /p/ responses to intended /b/ or /p/. In a similar fashion, the dotted and dashed lines show the results for the apical and velar categories. Each point in each graph is the average of 80 observations.[5]

### Place Identification by Following Vowel

Let us look first at the place scores from the analysis of all stimuli by following vowel (Figure 5).

Condition F123. The conversion of natural normal speech to synthetic parameters for (burst +) $F_1 + F_2 + F_3$ resulted in a loss of place information in specific cases. Nevertheless, place identification reached 93% overall. However, place identification was significantly lower before /i/ than before /a/, /u/ and /ɝ/ ($p \leq .006$, $p \leq .01$ and $p \leq .006$, two-tailed).

Condition F23. Eliminating $F_1$ to produce two-formant synthetic stimuli consisting of the (burst +) $F_2 + F_3$ lowered place identification scores a significant ($p \leq .01$) but small amount, to 90%. Place identification was still weaker before /i/ than before /a/ and /ɝ/ ($p \leq .026$ and $p \leq .018$, two tailed). Identification of bilabials was so low before /u/ that average place identification was as low before /u/ as before /i/. Even so, place of articulation generally remained highly identifiable in the absence of $F_1$.

Condition F2. Eliminating $F_1$ and $F_3$ to produce single formant stimuli consisting of the (burst +) $F_2$, yielded place identification scores significantly below those for (burst +) $F_2 + F_3$ ($p \leq .003$), at 72%. Place identification for bilabials and apicals averaged close to 90% before the classic experimental vowel /a/. Place identification was once again weaker before /i/ than before /a/, /u/ and /ɝ/ ($p \leq .006$, $p \leq .012$ and $p \leq .008$). Before /i/, place identification was lower for velar stops than for bilabial or apical stops ($p \leq .038$ or $p \leq .012$, two-tailed.

Condition F2-. Removing the bursts had the following effect on the place scores for $F_2$. Scores improved for the bilabials before /i/, compared with Condition F2 ($p \leq .028$, two-tailed). Overall place identification was essentially unchanged from Condition F2, at 73%.

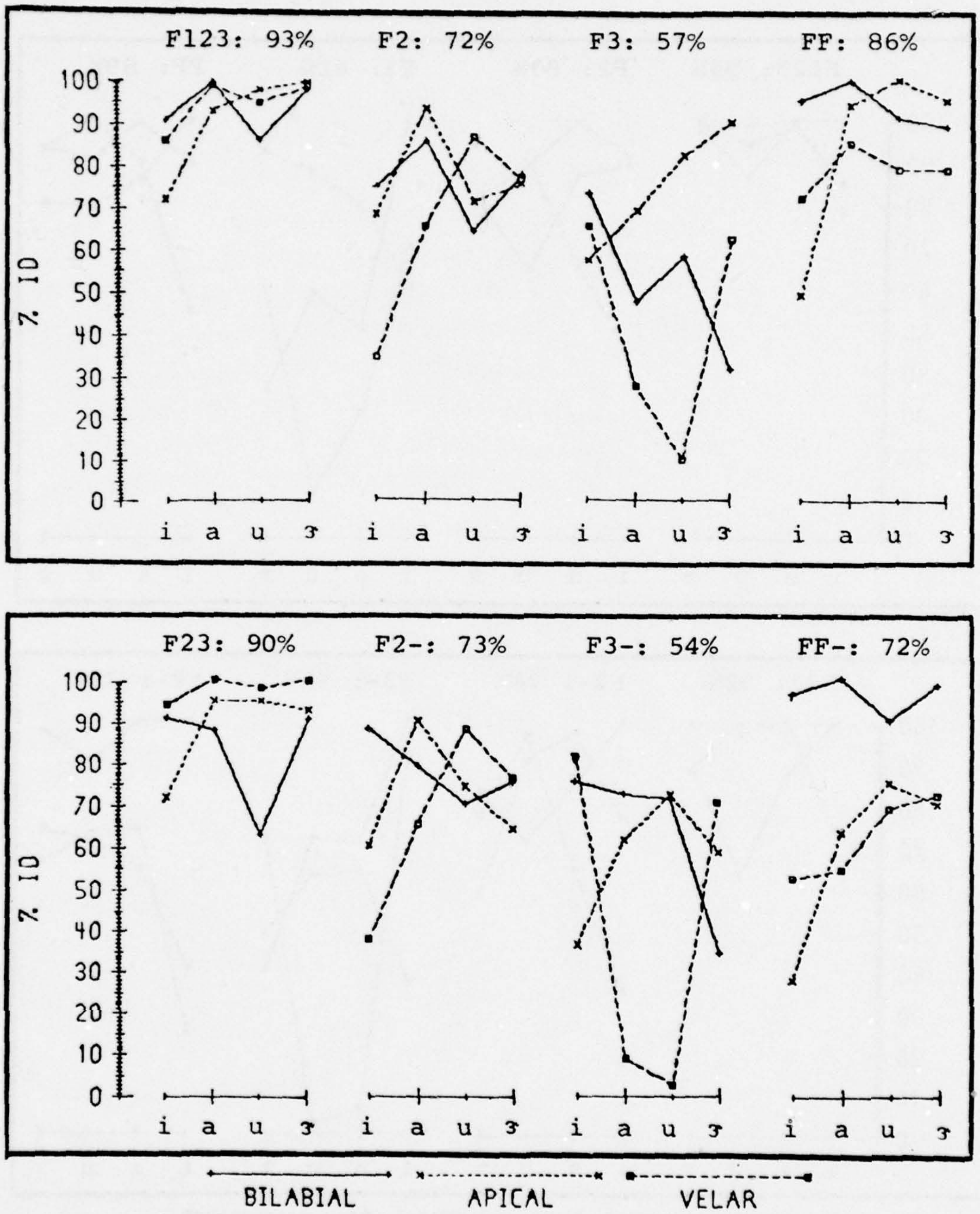


Figure 5: Place identification, by condition, from the analysis of all stimuli by following vowel.

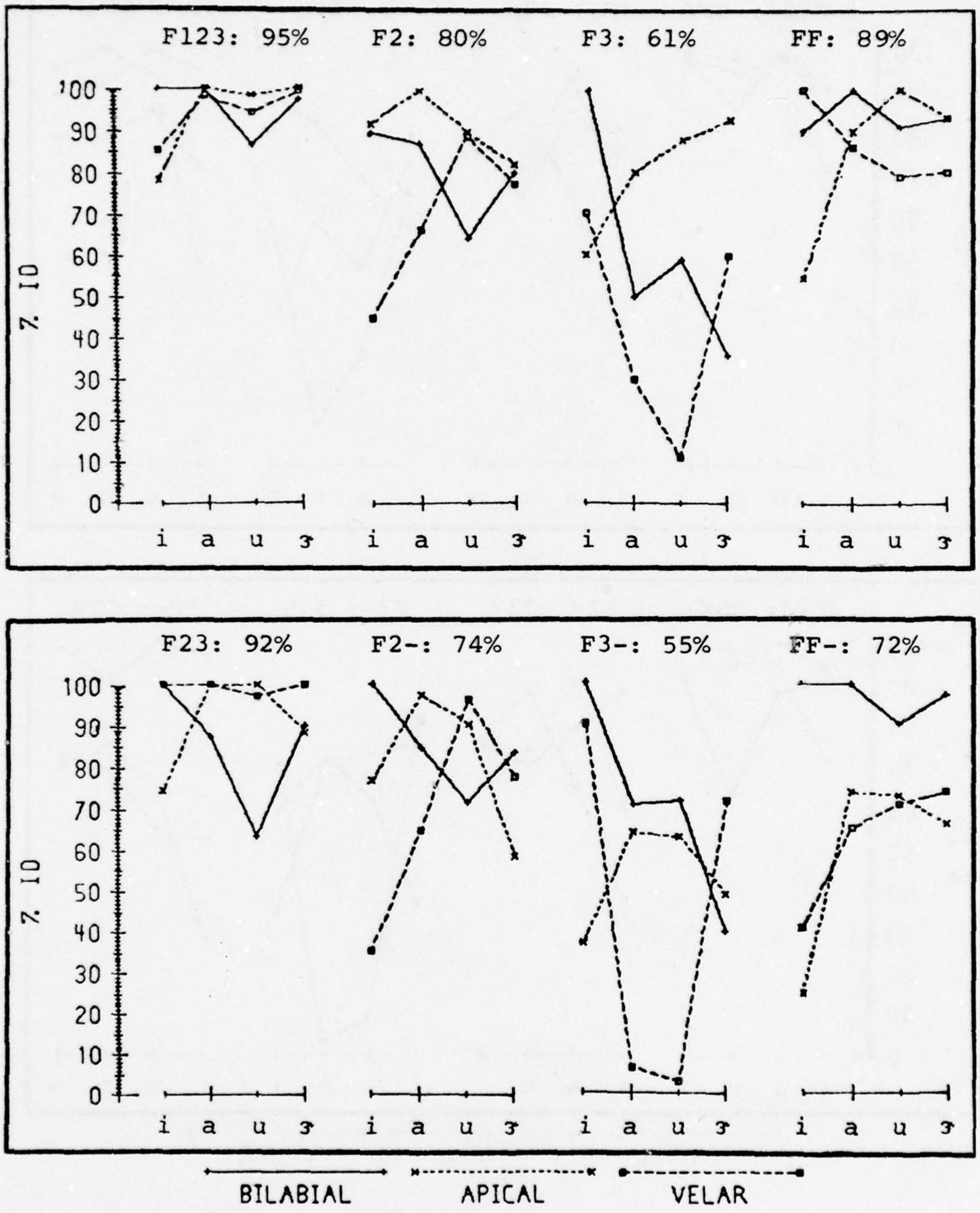


Figure 6: Place identification, by condition, from the analysis of the subset of stimuli synthesized with a burst, by following vowel.

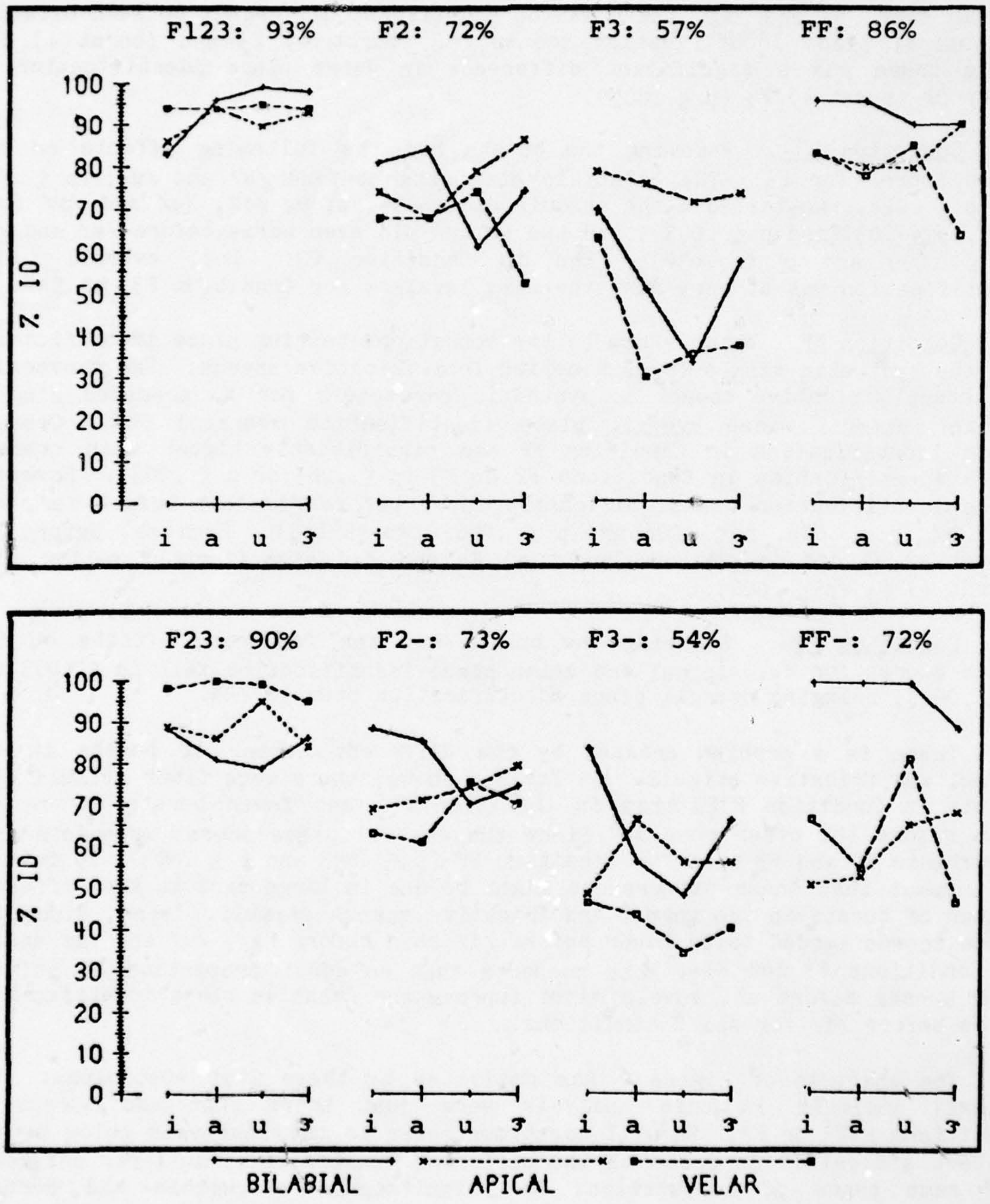


Figure 7: Place identification, by condition, from the analysis of all stimuli by preceding vowel.

*Condition F3*. Eliminating $F_1$ and $F_2$ to produce single formant stimuli consisting of the (burst +) $F_3$ yielded place scores significantly lower *overall* than for (burst +) $F_2$ ($p \leq .003$), at 57%. However, place scores in this condition were higher before /i/ and /ɝ/ than before /a/ or /u/ (before /i/: $p \leq .012$ or $p \leq .020$; before /ɝ/: $p \leq .008$ or $p \leq .020$, two-tailed). And in fact, *before /i/*, there was no significant difference between bilabial and apical place identification scores for (burst +) $F_2$ and (burst +) $F_3$, while there was a significant difference in velar place identification in favor of (burst +) $F_3$ ($p \leq .005$).

*Condition F3-*. Removing the bursts had the following effects on the place scores for $F_3$. The bilabials did better before /a/ and /u/, ($p \leq .008$ and $p \leq .028$, two-tailed), the apicals did worse before /i/, /a/ and /ɝ/ ($p \leq .005$, $p \leq .037$ and $p \leq .003$), and the velars did even worse before /a/ and /u/ ($p \leq .009$ and $p \leq .034$), than in Condition F3. But, overall place identification was at very much the same level as for Condition F3, at 54%.

*Condition FF*. We move now to the conditions testing place identification for the synthetic single formant copied from fricative speech. The conversion of natural fricative speech to synthetic parameters for $F_f$ produced single-formant stimuli whose *overall* place identification averaged 86%. Overall place identification in Condition FF was significantly higher than overall place identification in Conditions F2 or F3 ($p \leq .003$ or $p \leq .003$). However, place identification was significantly lower before /i/ than before /a/, /u/ and /ɝ/ ($p \leq .008$, $p \leq .006$ and $p \leq .006$, two-tailed). Even so, *before /i/* (burst +) $F_f$ did as well as (burst +) $F_3$ and did significantly better than (burst +) $F_2$ ($p \leq .006$).

*Condition FF-*. Removing the bursts had the following effects on the place scores for $F_f$. Apical and velar place identification fell ($p \leq .003$ and $p \leq .004$), bringing overall place identification down to 72%.

There is a problem created by the different number of bursts in the normal and fricative stimuli. As Table 4 shows, there were fewer stimuli with bursts in Condition F123 than in Condition FF, and fewer bursts before /i/ than before the other vowels.[6] Since the overall place scores were lower for Conditions F2 and F3 than for Condition FF ($p \leq .003$ and $p \leq .003$), it is easy to suspect that these differences might be due in large part to the different number of bursts in the normal and fricative speech stimuli. Also, since the place scores tended to be lower before /i/ than before /a/, /u/ and /ɝ/ except in Conditions F3 and F3-,[7] one suspects that an equal proportion of stimuli with bursts before all vowels might improve the relative place identification score before /i/ for all 8 conditions.

The analysis of Figure 6 was motivated by these last suspicions. The stimuli included in this analysis were just those that had bursts in Conditions F123 or FF. We will sometimes refer to this analysis below as the "subset analysis." In spite of the different numbers of stimuli per category, the same types of computations for significance of within- and across-condition differences were carried out. The different number of stimuli per category should be kept in mind when interpreting the probabilities reported below for this analysis.

NORMAL SPEECH BURSTS

|  | i | a | u | ɝ | ALL |
|---|---|---|---|---|---|
| b | 0 | 3 | 4 | 4 | 11 |
| p | 1 | 4 | 4 | 3 | 12 |
| d | 2 | 0 | 3 | 4 | 9 |
| t | 3 | 3 | 2 | 2 | 10 |
| g | 1 | 2 | 3 | 3 | 9 |
| k | 1 | 3 | 4 | 4 | 12 |
| ALL | 8 | 15 | 20 | 20 | 63 (65%) |

FRICATIVE SPEECH BURSTS

|  | i | a | u | ɝ | ALL |
|---|---|---|---|---|---|
| b | 2 | 4 | 4 | 4 | 14 |
| p | 3 | 4 | 4 | 4 | 15 |
| d | 0 | 4 | 4 | 4 | 12 |
| t | 3 | 4 | 4 | 4 | 15 |
| g | 2 | 4 | 4 | 1 | 11 |
| k | 3 | 2 | 3 | 4 | 12 |
| ALL | 13 | 22 | 23 | 21 | 79 (82%) |

Table 4: The number of bursts, by consonant, and following vowel, for the synthetic normal speech stimuli (top) and the synthetic fricative speech stimuli (bottom).

As Figure 6 begins to suggest, if we look at just those stimuli synthesized with a burst in Condition F123 or Condition FF, the place scores are still significantly higher for $F_f$ than for $F_2$ or $F_3$ ($p \leq .008$ or $p \leq .003$). This result does not indicate that the differences between overall place identification for these three conditions in the earlier analysis were largely due to the differing number of bursts. Overall place identification in Conditions F123, F2, F3 and FF is improved before /i/, but the small and variable number of stimuli makes it undesirable to attach any confidence to this finding. We should conclude only that the difference between the overall place scores for $F_f$, $F_2$ and $F_3$ in the earlier analysis is very probably not accounted for by the different number of bursts.

## Place Identification by Preceding Vowel

Let us look briefly at Figure 7. To compare the effects of preceding and following vowels thoroughly is beyond the scope of this study, but we said earlier that the intervocalic environment gives both closure and release information about the consonant. To the extent that this is so, it might be expected that effects of the changing phonetic context would be found if the data were analyzed in terms of the preceding vowel environment, averaging over the following vowels instead. Figure 7 presents the place identification scores from such an analysis by preceding vowel.

Overall, the preceding vowel environment appears to have had less effect on place perception than the following vowel environment: there are approximately 25% fewer significant within-condition differences (vowel to vowel, or place to place) in the analysis by preceding vowel environment than in the analysis by following vowel environment. A similarity between the preceding and following vowel environments, however, is that, on either side of the consonant, only one vowel yielded overall place identification that was as good for $F_3$ as for $F_2$, and that vowel was /i/. After /a/, /u/ and /ɝ/, $F_2$ provided better place identification than $F_3$ ($p \leq .004$, $p \leq .003$ and $p \leq .004$).

## IV. DISCUSSION

The high level of consonant place identification for the *natural fricative speech stimuli* in the preliminary experiment lends some credibility to the claim about the intelligibility of fricative speech. Of course, the perceptual result that the place of articulation of fricative speech stop consonants may be intelligible does not prove the physical front cavity hypothesis for normal speech. The physical hypothesis needs to be tested, for example, by an analysis of the energy in the vocal tract during various consonant articulations. Such a test could come from a study similar to that undertaken by Fant and Pauli (1974, pp. 121-132) for vowel articulations.

However, the results with the *synthetic single formant stimuli* did not always meet our predictions. Table 5 presents relevant significance test results from the analysis of all stimuli by following vowel. Differing results from the subset analysis are reported but a separate table is not given. In each row of Table 5, a predicted outcome can be compared with the observed outcome for each (bilabial, apical, velar) or all (total) places of articulation. Note that the predicted outcomes entered for the rows "without burst" are the same as those for the rows "with any burst" on the grounds that

CONTRIBUTION OF THE FORMANT

| CONDITIONS | VOWEL | PREDICTED | OBSERVED | | | |
|---|---|---|---|---|---|---|
| | | | BIL | APL | VEL | TOT |
| With any burst | before /i/ | FF = F3 | ><br>.009 | = | = | = |
| | | FF > F2 | ><br>.009 | <<br>.009 | ><br>.003 | ><br>.006 |
| | | F3 > F2 | = | = | ><br>.005 | = |
| | before /a/, /u/, /ɝ/ | FF > F3 | >>> | >>= | >>> | >>> |
| | | FF = F2 | >>> | =>> | =<= | >>> |
| | | F3 < F2 | <=< | <=> | <<< | <<< |
| Without burst | before /i/ | FF- = F3- | ><br>.006 | <<br>.029 | <<br>.003 | = |
| | | FF- > F2- | = | <<br>.008 | ><br>.018 | = |
| | | F3- > F2- | <<br>.014 | <<br>.033 | ><br>.004 | = |
| | before /a/, /u/, /ɝ/ | FF- > F3- | >>> | === | >>= | >>> |
| | | FF- = F2- | >>> | <== | <<= | ==> |
| | | F3- < F2- | ==< | <=< | <<= | <<< |

Table 5: The predicted versus observed outcome of the single-formant comparisons, from the analysis of all stimuli by following vowel. See further details in the text.

formant-cavity affiliations are not changed by the exclusion of the burst. The entries in the "observed" columns indicate either that there was no significant difference in place identification for the formants (=), or that the formant on the left in the predicted column for that row provided significantly (higher/lower) place identification scores than the formant on the right (>/<). The rows labelled "before /i/" give the outcome before the vowel /i/. The rows labelled "before /a/, /u/, /ɝ/" give the outcome, from left to right within each column, before the vowels /a/, /u/ and /ɝ/. Thus, there is one entry in every "observed" column in the "/i/" rows but three entries in every "observed" column in the "/a/, /u/, /ɝ/" rows.

Before /i/. As predicted, $F_f$ provided overall place scores equal to those for $F_3$. In the subset analysis, however, without the bursts, $F_f$ did worse overall than $F_3$ ($p \leq .004$). $F_f$ was also predicted to do better than $F_2$. With any bursts, $F_f$ did do better than $F_2$ ($p \leq .006$), overall. Without bursts, $F_f$ did as well as $F_2$, overall. In the subset analysis, however, with bursts, $F_f$ only did as well as $F_2$, overall, and without bursts, $F_f$ did worse than $F_2$ ($p \leq .018$, two-tailed), overall. Finally, $F_3$ was predicted to do better than $F_2$. With any bursts, $F_3$ did as well as $F_2$, overall. $F_3$ of velars did better than $F_2$ of velars, but $F_3$ of apicals and bilabials only did as well as $F_2$ of apicals and bilabials. Without bursts $F_3$ again did as well, overall, as $F_2$. In the subset analysis, with bursts, $F_3$ just barely did worse than $F_2$ ($p \leq .066$, two-tailed) overall.

Before /a/, /u/, and /ɝ/. As predicted, $F_f$ provided overall place scores higher than those for $F_3$ ($p \leq .003$ before each vowel). $F_f$ was also predicted to do as well as $F_2$. With any bursts, $F_f$ did better than $F_2$ ($p \leq .004$, $p \leq .003$ and $p \leq .014$), overall. Without bursts, $F_f$ did as well as $F_2$ before /a/ and /u/, and better than $F_2$ before /ɝ/ ($p \leq .04$), overall. In the subset analysis, without bursts, $F_f$ did as well as $F_2$ before all three vowels. Finally, as predicted, with or without bursts, $F_3$ did worse than $F_2$ ($p \leq .006$, $p \leq .003$ and $p \leq .004$) overall.

Thus, the results of the comparison of $F_f$ with $F_3$ or $F_2$ did not always turn out as predicted. $F_f$, the formant with the greatest expected front cavity affiliation overall, did produce the highest place identification scores overall of any of the single formants, as would have been predicted from the combination of formant-by-formant predictions. And, $F_3$ produced place identification scores as high as those for $F_f$ only before /i/, as predicted. But, whether $F_2$ did worse than $F_f$ or as well as $F_f$ depended not so much on the vowel context, as was expected, but on whether or not $F_f$ was synthesized with a burst. Often, the transition from the burst to the following formant was more continuous in Condition FF than in Conditions F2 and F3. Perhaps this gave the burst more perceptual weight in the fricative speech tests (see, e.g., Stevens & Blumstein, 1975).

As far as the results for $F_3$ compared to $F_2$ are concerned, our primary purpose was to ask whether or not $F_3$ did better than $F_2$ at providing place identification when it enjoyed the hypothesized closer affiliation with the front cavity. It appears that $F_3$ did as well as or better than $F_2$ at providing stop consonant place perception only for those stimuli in which the stop consonant occurred before a following /i/. This experimental result shows a change in the relative ability of $F_2$ and $F_3$ to provide place perception that is in the direction of the expected cavity affiliation change.

It does not, however, show as complete a change as was predicted: velar place identification was consistently better for $F_3$ than for $F_2$ before /i/, but bilabial and apical place identification scores were not significantly different for $F_3$ and $F_2$ before /i/.

From the point of view of the front cavity hypothesis, it seems clear why $F_3$ produced better *velar* place scores than $F_2$ before /i/. In the consonant release for the fricative speech velars before /i/, the bursts were close to 3000 Hz in front of a rather straight formant at about 3200 Hz. Similarly, in the normal speech velars before /i/ the bursts were at about 3000 Hz and were in front of an $F_3$ at about 3000 Hz. In these velar stimuli, then, the normal speech energy that was closest to the fricative speech energy appeared in and stayed in the $F_3$ region of the normal speech, and $F_3$ provided higher place identification scores than $F_2$.

It also seems reasonable, in hindsight, that $F_3$ might only do as well as $F_2$ in cueing *bilabial* place perception before /i/. In the consonant release for the fricative speech bilabials before /i/, the front cavity resonance rose (after any burst) from less than 2000 Hz to about 3000 Hz. In the normal speech bilabials before /i/, the consonant release also showed energy transitioning from less than 2000 Hz to about 3000 Hz. But in the normal speech case, the energy below 2000 Hz was part of $F_2$ while the energy at higher frequencies appeared as part of a rising transition in $F_3$. In these stimuli, then, the transitioning energy appeared first in the spectral region that corresponds to the $F_2$ region of normal speech, and subsequently passed into the $F_3$ region. According to the place identification scores, the part of this formant-crossing transition that was in the $F_2$ region cued place perception about as well as the part that was in the $F_3$ region.

But it is not clear why $F_3$ only did as well as $F_2$ at cueing *apical* place perception before /i/. In the consonant release for the fricative speech apicals before /i/, the bursts were at or above 3000 Hz, and the front cavity resonance rose from about 2500 Hz to about 3100 Hz. In the consonant release for the normal speech apicals before /i/, $F_3$ moved from about 2700 Hz to 3000 Hz, while $F_2$ was flat or slightly rising at about 2200 Hz. In other words, $F_f$ looked more like $F_3$, but $F_3$ did not do better than $F_2$. In addition, apical place perception was generally lower before /i/ than before /a/, /u/, or /ɝ/. There were short and sometimes rising transitions, and fewer bursts for the apicals before /i/. This suggests a possible causal relationship between short, rising transitions and lower apical scores, regardless of the front cavity affiliations.[8] We shall have to find out how much these apical place scores are improved by higher quality synthesis (better formant frequency, bandwidth, and amplitude control), before trying to interpret the results relative to the front cavity hypothesis.[9]

Despite possible shortcomings in the synthesis of the apicals before /i/, it seems clear that perceptually important differences in the role of $F_2$ and $F_3$ were found across the bilabial, apical and velar stimuli before this single vowel. These specific differences are consistent with our understanding of the manifestation of the front cavity resonance, if not with the averaged predictions that were expressed entirely in terms of the expected front-cavity affiliation with $F_3$ of /i/. The net result appears to be that the front-cavity affiliation and the ability of $F_2$ and $F_3$ to provide place identification have varied as a function of both the vowel context and the consonantal place of articulation.

## Contribution of the Bursts

Table 6 presents a selection of significance test results that is helpful in summarizing the contribution of the bursts. These are the results of the comparisons of place scores for each formant with and without the bursts, from both analyses by following vowel: the results from the analysis of all stimuli are presented in the top half of the table; those from the analysis of the subset of stimuli synthesized with a burst are presented in the bottom half of the table. Since the predictions are not different for any formant before /i/, the results before /i/, /a/, /u/ and /ɝ/ are now presented on the same line, from left to right within each column. The interpretation of the equality or inequality signs is the same as in Table 5.

---

Table 6: The predicted versus observed outcome of the comparisons of single-formant stimuli with and without bursts, from the analysis of all stimuli by following vowel (top), and from the analysis of the subset of stimuli synthesized with a burst (bottom). See further details in the text.

CONTRIBUTION OF THE BURST

| ANALYSIS | PREDICTED | OBSERVED | | | |
|---|---|---|---|---|---|
| | | BIL | APL | VEL | TOT |
| ALL STIMULI | FF > FF- | ===< | >>>> | >>>= | >>>> |
| | F2 > F2- | <=== | ==== | ==== | ==== |
| | F3 > F3- | =<<= | >>=> | =>>= | =>>= |
| SUBSET WITH BURST | FF > FF- | ==== | >=>> | >>== | >>>> |
| | F2 > F2- | ==== | >==> | ==<= | >==> |
| | F3 > F3- | =<<= | >>>> | =>>= | >==> |

---

The predictions in Table 6 simply say that, overall, each formant should do better with the burst than without, on the grounds that the burst provides additional information for place perception. However, persons familiar with the acoustic structure of stop consonants will recognize that there should be exceptions to this general prediction. For example, if the burst of a /ku/ is synthesized with a single spectral peak at the frequency of $F_2$, then $F_3$ of that stimulus may sound more like it had a bilabial stop when the burst is included. Or, if the burst of /pi/ is synthesized with a single peak near the frequency of $F_2$, then $F_2$ of that stimulus may sound more like it had a velar stop when the burst is included. Such exceptions to the general prediction remind us that the burst is not usually heard in the environment of a single formant. But at the same time they suggest that, in natural speech, the burst is not primarily interpreted with respect to that particular formant.

A first observation about the effect of the bursts is that the overall place scores were never worse with the bursts. This result is consistent with the general finding that bursts do provide information for place perception.

Second, however, there were several specific place scores that were worse with the bursts, and all but one occurred with intended bilabials, where: with the bursts before /ɝ/, $F_f$ was more often perceived as a velar; with the bursts before /i/, $F_2$ was more often perceived as a velar; with the bursts before /a/, $F_3$ was more often perceived as an apical or a velar; and with the bursts before /u/, $F_3$ was more often perceived as an apical. In addition, in the subset analysis of intended velars, the $F_2$ stimuli with bursts before /u/ were more often perceived as bilabials.

Since the bilabial bursts tended to be lower in amplitude than the apical or velar bursts, there is a possibility that they suffered disproportionately from the restricted dynamic range for synthesis. (See again the comment in section II.) Note, however, that overall place identification for the bilabial stimuli was equal to or better than that for the apical and velar stimuli in Conditions F123 and FF.

If the bilabial burst was as satisfactorily synthesized as the apical and velar bursts, then these results would suggest that the bilabial burst does not make the same contribution to bilabial place perception that the apical and velar bursts do to perception of their respective places of articulation. Under the front cavity hypothesis the differential contribution of the bilabial burst may be explained by the lack of a well-defined cavity in front of the bilabial constriction. From this point of view, the burst had greater place cue value when it could be interpreted as part of the same front cavity affiliated energy that was an important cue for the formants that followed. This equivalence of the place information in bursts and formants affiliated with the front cavity would help explain the paradoxical situation first mentioned by Halle, Hughes, and Radley (1957), where "what appear to be the two most disparate acoustical phenomena, formant movements and bursts of sound, are perceptually equated."

Third, the bursts made more of a contribution to place perception for $F_f$ than for $F_3$, and for $F_3$ than for $F_2$. Overall place identification for $F_f$ improved with the bursts before every vowel, as a result of the generally improved scores for apicals and velars with the bursts. Overall place identification for $F_3$ improved with the bursts for half of the vowels, with apical and velar scores again the most affected. Overall place identification for $F_2$, however, showed no improvement with the bursts in the analysis of all stimuli, and improvement only as a result of including the apical bursts in the subset analysis. If reliable, these results suggest that the ability of $F_3$ to cue place perception might have been underestimated compared to that for $F_2$, if our stimuli had not included bursts.

## Relationship to Other Hypotheses

The results of this experiment can be related to other hypotheses about stop consonant place perception that have recently been under investigation.

Klatt and Shattuck (1975, pp. 293-302) reported, as have a number of other researchers (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, pp. 13-50), that frequency transitions in the second formant region are of greater perceptual importance than frequency transitions in the third formant region, for stimuli with no burst. Those results appear generally to be true, but we have found evidence of a change in the relative importance of $F_2$ and $F_3$ before the vowel /i/. This one change is not inconsistent with the view that $F_2$ is usually more important than $F_3$. On the other hand, the change does not appear as an exception under a view of place perception that emphasizes the importance of the front cavity, rather than one that emphasizes a particular numbered formant.

Another relevant paper is Stevens and Blumstein (1975), where it was suggested that context-dependent formant transitions are secondary cues to consonant place of articulation. Our results indicate that the cue value of a given formant may vary a great deal from one articulatory configuration to another. For example, there was 35% place identification for $F_2$ of velars before /i/, but 90% place identification for $F_2$ of velars before /u/. This kind of variability may well be consistent with the idea that formant transitions do not make a constant contribution to place perception. However, these large variations still need an explanation. Our hypothesis would be that formant transitions, like the bursts, derive much cue value from close affiliation with the front cavity, which affects bursts, frication, aspiration, and voiced transitions from stop consonants.

Finally, it appears that spectra generated from the output of a cochlear model (Carlson, Fant, & Granstrom, 1975, pp. 55-82) have tended to show a single component prominence in place of the detailed upper formant structure, when the input to the model was either a close or an open vowel. The front cavity has an important effect on the frequency and shape of this prominence, and in cases of extreme vocal tract constriction, its own fundamental resonance lies in this region of upper spectrum energy. The result is that a formant close to the frequency of the front cavity resonance tends to be relatively loud. There is in addition some evidence that the loudest formants are the most important ones for perception of vowel color (see, e.g., the weighted averages of upper formants in Delattre et al., 1952, or the work of Carlson, Granstrom, & Fant, 1970, on an equivalent upper formant, "$F_2$ prime," for vowels). Consequently, it seems possible that place perception could be viewed as the interpretation of a loud component in the speech spectrum as a contribution specifically of the front cavity. What happens when the front cavity is not well defined? A person who is perceiving speech may simply interpret the loudest prominence as the indication of what the front cavity is doing. Interestingly, the fricative speech /ə/ is not located near 500 Hz, as it would be if it were simulating the true fundamental quarter-wave resonance of a uniform tube of the size of the vocal tract. Instead, it is located near 1500 Hz, i.e., in the region of $F_2$ of the uniform tube, and it sounds not unlike /ə/.

## V. CONCLUSION

Our purpose in this study has been to test the hypothesis that the individual formants of normal and fricative speech most closely affiliated with the front cavity are also most important for the perception of stop consonant place of articulation. Three of our findings tend to support this front cavity hypothesis. First, the front cavity resonance from fricative speech, $F_f$, provided the highest overall place identification scores of any of the single formants. Second, $F_2$ and $F_3$ from normal speech showed a change in their ability to provide place perception that is in the direction of the expected formant-cavity affiliation changes. Specifically, $F_3$ did as well as, or better than $F_2$ at providing stop consonant place perception only before the vowel /i/. Third, the stop burst seemed to be an important place cue only for the apical and velar stops, where it can be thought of as the initial manifestation of the front cavity resonance. A fourth finding suggested that accurate place identification scores can be obtained for the $F_2$ and $F_3$ of normal speech even if they are not the second and third formants in the spectrum: (burst +) $F_2$ + $F_3$ provided 90% place identification without $F_1$.

Such findings increase our confidence that the acoustic information that is most closely affiliated with the front cavity resonance is also the most important for the perception of the phonetic dimension of place of articulation, even though that affiliation is sometimes with different formants by number. Further experimental results are needed from speech synthesis (e.g., to settle questions about the apicals before /i/), and from speech analysis (e.g., for physical confirmation of the spatial distribution of energies in the vocal tract). But in the meantime, this experiment has provided some support for the perceptual hypothesis that the front cavity resonance can be used as an articulatory reference to explain what appears in the acoustic signal to be a multiplicity of cues to place perception.

## VI. REFERENCES

Carlson, R., Fant, G., & Granstrom, B. Two-formant models, pitch and vowel perception. In G. Fant & M. A. A. Tatham (Eds.), _Auditory analysis and perception of speech_. London: Academic, 1975.

Carlson, R., Granstrom, B., & Fant, G. Some studies concerning perception of isolated vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm), 1970, _STL-QPSR_ _2-3_, 19-35.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. _Journal of the Acoustical Society of America_, 1952, _24_, 597-606.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. An experimental study of the acoustic determinants of vowel color; observa-

observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, 1952, *8*, 195-210.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop consonant recognition: Release bursts and formant transitions as functionally-equivalent, context-dependent cues. *Haskins Laboratories Status Report on Speech Research*, 1976, *SR-47*, 1-27.

Epstein, R. A transistorized formant-type synthesizer. *Haskins Laboratories Status Report on Speech Research*, 1965, *SR-1*, 7.1.

Fant, C. G. M. *Acoustic theory of speech production* (2nd ed.). 's Gravenhage: Mouton, 1970. (Originally published, 1960.)

Fant, C. G. M. Auditory patterns of speech. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*. Cambridge, Mass.: MIT Press, 1967a.

Fant, C. G. M. Stops in CV syllables. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm), 1967b, *STL-QPSR 4/69*, 1-25.

Fant, C. G. M., & Pauli, S. Spatial characteristics of vocal-tract resonance modes. In *Preprints of the speech communications seminar* (Vol. II). Stockholm: 1974.

Fischer-Jorgensen, E. Acoustic analysis of stop consonants. *Miscellanea Phonetica*, 1954, *2*, 42-59.

Halle, M., Hughes, G. W., & Radley, J.-P. A. Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, 1957, *29*, 107-116.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. Effect of third-formant transitions on the perception of the voiced stop consonants. *Journal of the Acoustical Society of America*, 1958, *30*, 122-126.

Heinz, J. M., & Stevens, K. N. On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 1961, *33*, 589-596.

Klatt, D. H., & Shattuck, S. R. Perception of brief stimuli that resemble rapid formant transitions. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech*. London: Academic Press, 1975.

Kuhn, G. M. On the front cavity resonance and its possible role in speech perception. *Journal of the Acoustical Society of America*, 1975, *58*, 578-585.

Kuhn, G. M., & McGuire, R. M. Results of a VCV spectrogram reading experiment. *Haskins Laboratories Status Report on Speech Research*, 1974, *SR-40*, 67-80.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. In E. E. David & P. B. Denes (Eds.), *On human communication: A unified view*. New York: McGraw-Hill, 1967.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497-516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 1954, *68*, 1-13.

Ohman, S. E. G. Coarticulation in VCV utterances: A spectrographic analysis. *Journal of the Acoustical Society of America*, 1966, *39*, 151-168.

Peterson, G. E., & Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175-184.

Siegel, S. *Non-parametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Stevens, K. N. Acoustic correlates of place of articulation for stop and fricative consonants. MIT Research Laboratory of Electronics Quarterly Progress Report, 1968, *QPR 89*, 199-205.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory and acoustical data. In E. E. David & P. B. Denes (Eds.), *On human communication: A unified view*. New York: McGraw-Hill, 1972.

Stevens, K. N. Potential role of property detectors in the perception of consonants. MIT Research Laboratory of Electronics Quarterly Progress Report, 1973, *QPR 110*, 155-168.

Stevens, K. N., & Blumstein, S. Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, 1975, *3*, 215-233.

Stevens, K. N., & House, A. S. Studies of formant transitions using a vocal-tract analog. *Journal of the Acoustical Society of America*, 1956, *28*, 578-585.

Stevens, K. N., & House, A. S. An acoustical theory of vowel production and some of its implications. *Journal of Speech and Hearing Research*, 1961, *4*, 303-320.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1945, *1*, 80-83.

## FOOTNOTES

[1] We were referring to the fundamental quarter-wave resonance of the cavity behind the mouth opening, for articulations in which the mouth is open and tongue constriction is extreme. See Fant and Pauli (1974) for a more general treatment of formant-cavity affiliations in terms of the total reactive energy (kinetic plus potential) found in a localized region of the vocal tract.

[2] Since the amplitudes for the successive channels are estimated sequentially, from low frequency to high, while the waveform in the analyzer's recirculating memory is continuously updated, the amplitude estimates for the high end of the frequency scan correspond to slightly later moments in time than those at the lower end. However, for the purpose of estimating synthesis parameters, each frequency scan was treated as though it represented exactly one moment in time.

[3] Using the computer, we could draw a line under the desired spectral peak, and request that the exact frequency of the next higher peak be used as the specified formant.

[4] One important case was the $F_2$ of syllables with /i/, which often needed to be lower in amplitude than the $F_3$ of the same syllables. Formant level is usually inversely related to formant number (Peterson & Barney, 1952; Stevens & House, 1961; Klatt & Shattuck, 1975). But spectrograms showing the opposite relationship for $F_2$ and $F_3$ of /i/ are found in the literature (Fant, 1960, 1967b, pp. 111-125; Kuhn, 1975).

[5] Graphs of mean consonant identification by voicing category showed a large advantage for the voiced stops in Condition F23, the first experimental condition with aperiodic excitation. In this same condition there was also a strong response bias for the voiced stop symbols. No other condition showed this kind of difference between the voicing categories.

[6] The average duration of the bursts in both the synthetic normal speech and the synthetic fricative speech stimuli was 1.59 time frames (about 20 msec).

[7] In Conditions F123, F23, F2, F2-, FF and FF-, overall place identification was significantly lower before /i/ than before the other vowels ($p \leq .036$, two-tailed, in each case), except before /u/ in Condition F23, where there was no significant difference between the two vowel environments. In Conditions F3 and F3-, overall place identification was significantly higher before /i/ than before the other vowels ($p \leq .010$ in each case), except before /ɝ/ in Condition F3, where there was no significant difference between the two vowel environments.

[8] Similarly, a lower peak percentage of apical responses was obtained for rising transitions in Cooper, Delattre, Liberman, Borst, and Gerstman (1952). Note also that rising formant transitions observed for the apicals before /i/ are an exception to the rule that $F_2$, $F_3$ and $F_4$ transitions for non-retroflex apical consonants tend to fall (Stevens & Blumstein, 1975).

[9] In the consonant release of the fricative speech stimuli with an apical stop before a back vowel, the front cavity resonance started out at about 3000 Hz and then fell below 2000 Hz. In the consonant release of the normal speech stimuli with an apical stop before a back vowel, there was a transition from $F_3$ to $F_2$ that covered the same range. In the normal speech stimuli, this transition tended to produce a longer and longer fall in $F_3$ as the $F_3$ "target" became lower and lower for the vowels /a/, /u/, and /ɝ/. Note that in Figure 5, (burst +) $F_3$ produced better and better apical place identification across these same three vowels. These improving scores for $F_3$ of apicals are consistent with the cue value of long transitions often hypothesized in spectrographic analyses of speech (Fischer-Jorgensen, 1954; Dorman, Studdert-Kennedy, & Raphael, 1976).

# VOWEL DURATION CHANGE AND ITS UNDERLYING PHYSIOLOGICAL MECHANISMS*

Katherine S. Harris+

Abstract. Two explanations have been proposed for the relationship between vowel target formant frequency and articulatory stress. The first, the "extra energy" hypothesis, suggests that stressing is accompanied by larger signals to the articulators, so that stressed syllables are longer and have more extreme formant values. The second, the "undershoot" hypothesis, suggests that the signals sent to the articulators are of constant magnitude, but that changes in timing result in differences in formant frequency. This view leads to a prediction that the relationship between target formant frequency and duration is fixed, whatever the cause of the duration variation. Acoustic and electromyographic measures were made of productions of nonsense syllables with varying stress and speaking rate, by three speakers. Results fail to support the undershoot hypothesis, since syllable duration and vowel target frequency are independent. While speaking rate variations are accomplished in a different manner by the three speakers, the "extra energy" model for stressing seems to be supported.

## INTRODUCTION

The central problem of speech production research has been to explain allophonic variation; that is, the effect of various factors on the articulatory manifestation of a given phone. Traditionally, it has been common to separate context effects into two classes -- the so-called "coarticulation" and "timing" effects -- although the theoretical problems are the same for both classes. It is recognized also that, in general principle, some allophonic variations are to be considered either idiosyncratic or language specific, while others may be considered to arise from general properties of the motor organization of the articulatory system. It is with this last, poorly defined, class that most studies have been concerned.

*In press, Language and Speech. (Earlier versions of this paper were presented at the 50th Meeting of the American Speech and Hearing Association, Washington, D. C., November 1975, and at the 94th meeting of the Acoustical Society of America, Miami, Florida, December, 1977.)

+Also, The Graduate School, The City University of New York

Acknowledgment: This research was supported by NIH grants DE-01774 and NS-13617 to Haskins Laboratories. The techniques for making the acoustic measurements were developed by Thomas Baer; the difficult task of analyzing the data was accomplished by Betty Tuller.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

One of the best studied allophonic effects is the effect of syllable stress on vowel production. It has been shown that stressed vowels are commonly both more intense and longer than their unstressed counterparts. In addition, stressed and unstressed vowels differ in vowel quality, in .that unstressed vowels tend to be neutralized (Lindblom, 1963). The vowel quality difference is sufficiently substantial that it has been shown to cue the perception of stress in disyllables (Fry, 1964). This sort of quality difference has been shown for a number of languages (Lehiste, 1970).

Two models have been proposed for the effect of stress on vowel color. The first might be called an "extra energy" model. While the details of the model are not worked out at a physiological level, the general idea is that extra energy is applied to the stressed vowel, with the result that it lasts longer, and the signals to the articulators are a little larger, so that the vowel is further from a neutral vocal tract position (Ohman, 1967; Jones, 1932).

A second model might be called the "undershoot model" (Lindblom, op. cit). In Lindblom's model, the difference between target formant values for the vowels of differing duration is a consequence of the change in duration itself. A vowel is specified in the nervous system by a set of signals. When these signals are sent to the articulators, they result in a given vocal tract shape, the target, unless the signals for a subsequent phone arrive at the articulators too soon, so that the target is not attained because the path of articulatory movement is deflected towards the new target.

These two theories make different predictions about events at the acoustic and physiological levels. For the "undershoot" theory, at the acoustic level, the effect of any change of context on the relationship between duration and formant target frequency is the same, so that the effects of changing stress and speaking rate, for example, are identical. A given decrease in duration will be accompanied by a fixed undershoot. At the control signal level, the size of the signals to the articulators should be constant, as stress or speaking rate is varied. For the "extra energy" hypothesis, vowel duration and target frequency are separable, and stressed syllables should show more extreme formant values than unstressed syllables.

The experiment to be described was developed to test these theories, and beyond that, to gain insight into the speech timing mechanism, by studying the effect of stress and speaking rate on simple nonsense syllables.

## METHODS

The speakers were three adults, two females (KSH and FBB) and one male (LJR), all native speakers of American English, and personnel of Haskins Laboratories.

The speech sample was the four-syllable nonsense words /əpipipə/ and /əpipibə/, with stress placed on either the second or third syllable. Subjects read semi-random lists of these four "words" at two self-selected speaking rates, "slow" and "fast." Although 25 repetitions were produced of each utterance, later processing failures reduced the lists to 24 repetitions for LJR and 20 and 21, respectively, for KSH and FBB. Acoustic recordings were made, as well as electromyographic recordings from the genioglossus muscle. Since the genioglossus bunches the main body of the tongue, and brings it forward (Raphael & Bell-Berti, 1975; Smith 1970), we might expect greater activity from the muscle as fronting increases. In order to ensure at least one successful recording, two electrodes were inserted into the muscle for KSH and FBB, and three for LJR. Since all insertions were successful, we selected those recordings which appeared, on preliminary inspection, to be most stable, for further analysis. The electrode preparation and insertion technique has been reported in detail elsewhere (Hirose, 1971).

All acoustic measurements were made on an interactive computer system at Haskins Laboratories. Duration measures were made on the waveform; the duration for each syllable represents the duration of closure, burst, aspiration and voicing for the central two syllables in the nonsense word, indicated as the first and second syllables below. Measurements of $F_2$ and $F_3$ peak frequency were made after spectrographic transformation, although as the measures of $F_2$ frequency were so closely parallel to those of $F_3$ frequency, they will not be reported. Since low frequency room noise was recorded during the experiment, reliable measurement of $F_1$ frequency was not feasible.

The EMG signals were rectified, filtered and averaged using the Haskins Laboratories system, as previously described (Port, 1971; Kewley-Port, 1973). Peak EMG activity was measured for each syllable, as a crude indication of overall muscle activity. Some typical averaged interference patterns are shown in Figure 1.

## RESULTS

Results of the acoustic measurements are shown in Figure 2. Each panel shows the eight data points for $F_3$ peak frequency as a function of duration, for /p/ or /b/ syllables, for a single speaker. Each point is labelled as to whether it was produced in a first or second syllable, as to the stressed or unstressed character of the syllable, and the speaking rate condition. In the top set of six panels, lines are drawn between syllables which show a minimal contrast in speaking rate. If the "undershoot" hypothesis were supported, we would expect all lines to slope upward and to the right; that is, "slow" syllables would be longer in duration, indicating that the subject was following instructions, and that the fast syllables would have lower $F_3$ frequency values, indicating undershoot at fast speaking rates. In fact, the proposed pattern is followed, for two subjects, LJR and FBB, but is not for the third, KSH.

In the lower set of six panels, lines are drawn between points which show a minimal stress contrast. For contrasts of stress, lines connecting corresponding syllable pairs slope upward and to the right for all three subjects. Furthermore, in those cases where there is an overlap of duration values from different syllables, there is a tendency for stressed syllables to lie at higher values of $F_3$ peak frequency than unstressed syllables.

There are two minor variables which might have shown some systematic pattern. Since the nonsense words were produced in citation form, somewhat longer durations might be expected for the second syllable; furthermore, longer durations should be expected for those syllables terminating in /b/. A comparison of the twelve relevant cases shows that while all $S_2$ /b/ syllables are longer than $S_2$ /p/ syllables, $F_3$ frequency values are higher for /b/ for only five of the twelve comparisons, indicating no systematic relationship between the two variables. In short, speaking rate variation seems to be quite different from stress variation in its effects on vowel target; and individuals differ in how they accomplish speaking rate variation. Furthermore, the relatively small variations in duration which accompany the effect of terminal voicing and phonetic environment do not seem to be accompanied by systematic changes in $F_3$ frequency.

Figure 3 shows peak EMG activity as a function of acoustic duration, in an analagous form of presentation to that of Figure 1. It is clear that the second Lindblom hypothesis is not supported; peak activity varies substantially for stress, for most comparisons, for all three subjects; for speaking rate, there are no consistent effects on peak activity.

As to the effects of the minor variables of syllable position and terminal consonant, out of twelve possible comparisons, peak activity for /b/-syllables is greater than that for /p/-syllables in six. Thus, the overall conclusions from a study of peak EMG activity is that signals to the articulators are not of constant size; stressed syllables show greater activity than do unstressed syllables. Other variables do not show consistent effects.

## DISCUSSION

The results of the experiment lead to the general conclusion that the mechanisms for the control of speaking rate and duration in vowel articulation are disjunct, a conclusion which finds some support in the literature.

As to the acoustic result, the experiment closest to that reported above is that of Gay (1978). He concluded that speaking rate and lexical stress are controlled by different mechanisms, on the basis of results which show variations of vowel target frequency as a function of stress, but no variation as a function of speaking rate. Our results are identical to his for stress, but at variance with respect to speaking rate.

There are some studies of articulator movement under conditions of varying stress and speaking rate which, although no acoustic measures are reported, are consonant with those reported here. Kent and Netsell (1971), in a cineradiographic study, examined articulator position as a function of stress. Most of their observations show that under conditions of contrastive stress, the articulators move further from neutral position, and there is, in addition, some evidence for increased articulatory velocity with increased stress.

Using a technique identical to that of Kent and Netsell, Kuehn and Moll (1976) examined articulator displacement for vowels under conditions of variable speaking rate. They found that, as speaking rate increased, some speakers decreased articulator target position, while others increased articulator velocity to reach the target in a reduced amount of time.

Finally, in a study of single motor units, in the anterior belly of the digastric muscle, Sussman and MacNeilage (1978) observed a pattern of recruitment and discharge reorganization in emphatic stress, characterized by, among other things, changes in motor unit discharge rate and recruitment of additional motor units. They conclude that their findings support the Ohman (1967) notion of "an instantaneous addition of a quantum of physiological energy underlying stressed productions."

It can be concluded, then, that speakers articulate using independent controls of the duration and magnitude of movement, in generating allophonic variations in varying stress and speaking rate. There is some acoustic evidence (Nord, 1974) that the mechanism associated with terminal lengthening may lead to allophones with still a different relationship between duration and target frequency. Terminal lengthening may be particularly important in conveying syntactic information about sentence structure to the listener (Klatt, 1976; Cooper, 1976).

There is a possible reason, from a perceptual point of view, why these independent controls are necessary. Klatt (1976) has pointed out that, for a listener faced with the acoustic representation of a sentence, there is information encoded in duration about both the suprasegmental and segmental structure of the sentence. He suggested that if the listener already understood the sentence, he could interpret the durational variations, but he must use the durational information to decode the message. However, if durational variation is accomplished so that there are different relations among duration, target formant frequencies and transition velocity, for, for example, stressing and clause terminal lengthening, some of the ambiguities in the message may be resolvable.

It has recently been shown by Strange, Verbrugge, Shankweiler and Edman (1976) that vowel identification is aided by consonant context; presumably, as we understand movement dynamics better, we will better understand the way in which the listener makes use of the acoustic counterpart of the articulatory act.

## REFERENCES

Cooper, W. E. Syntactic control of timing in speech production. Journal of Phonetics, 1976, 4, 151-171.

Fry, D. B. The dependence of stress judgments on vowel formant structure. In Proceedings of the 5th International Congress of Phonetic Sciences, Munster, pp. 306-311. Basel: S. Karger, 1964.

Gay, T. Effect of speaking rate on vowel formant movements. Journal of the Acoustical Society of America, 1978, 63, 223-230.

Hirose, H. Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research, 1971, SR-25/26, 73-86.

Jones, D. An outline of English phonetics (3rd ed.). New York: Dutton, 1932.

Kent, R., & Netsell, R. Effects of stress contrasts on certain articulatory parameters. Phonetica, 1971, 24, 23-44.

Kewley-Port, D. Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research, 1973, SR-33, 173-

183.

Klatt, D. The linguistic uses of segment duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Kuehn, D. P., & Moll, K. L. A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 1976, *4*, 303-320.

Lehiste, I. *Suprasegmentals*. Cambridge: M.I.T. Press, 1970.

Lindblom, B. E. F. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, *35*, 1773-1781.

Nord, L. Vowel reduction - centralization or contextual assimilation? *Preprints of the Speech Communication Seminar, Stockholm, August 1-3*, 1974, *V2*, 149-154.

Ohman, S. E. G. Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory, Royal Institute of Technology, Quarterly Progress and Status Report*, 1967, *2-3*, 20-54.

Port, D. K. The EMG data system. *Haskins Laboratories Status Report on Speech Research*, 1971, *SR-25/26*, 67-72.

Raphael, L. J., & Bell-Berti, F. Tongue musculature and the feature of tension in English vowels. *Phonetica*, 1975, *32*, 61-73.

Smith, T. *A Phonetic Study of the Function of the Extrinsic Tongue Muscles*. Unpublished Ph. D. dissertation, U. C. L. A., 1970.

Strange, W., Verbrugge, R. R., Shankweiler, D., & Edman, T. Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 1976, *60*, 213-244.

Sussman, H., & MacNeilage, P. F. Motor unit correlates of stress: Preliminary observations. *Journal of the Acoustical Society of America*, 1978, *64*, 338-340.

# FIGURE LEGENDS

Figure 1. Averaged EMG activity for the genioglossus muscle for two speakers. Stress contrasts are shown at the left; speaking rate contrasts are shown at the right.

Figure 2. Peak $F_3$ frequency plotted against syllable duration, for three subjects. Values for utterances whose final consonant is /p/ are plotted in the first and third rows; those for utterances whose final consonant is /b/ arc plotted in the second and fourth rows. Points representing minimal contrasts in speaking rate are connected in the upper six panels; points representing minimal stress contrasts are connected in the lower six panels. Stressed syllables are indicated with upper case letters; unstressed syllables are indicated with lower case letters. Values for first syllables are subscripted with "1"; those for second syllables, with "2." Values for fast speaking rate are indicated with "F"; those for slow speaking rate are indicated with "S."

Figure 3. Peak EMG activity (in microvolts) plotted against syllable duration, for three subjects. Values for utterances whose final consonant is /p/ are plotted in the first and third rows; those for utterances whose final consonant is /b/ are plotted in the second and fourth rows. Points representing minimal contrasts in speaking rate are connected in the upper six panels; points representing minimal stress contrasts are connected in the lower six panels. Stressed syllables are indicated with upper case letters; unstressed syllables are indicated with lower case letters. Values for first syllables are subscripted with "1"; those for second syllables, with "2." Values for fast speaking rate are indicated with "F"; those for slow speaking rate are indicated with "S."

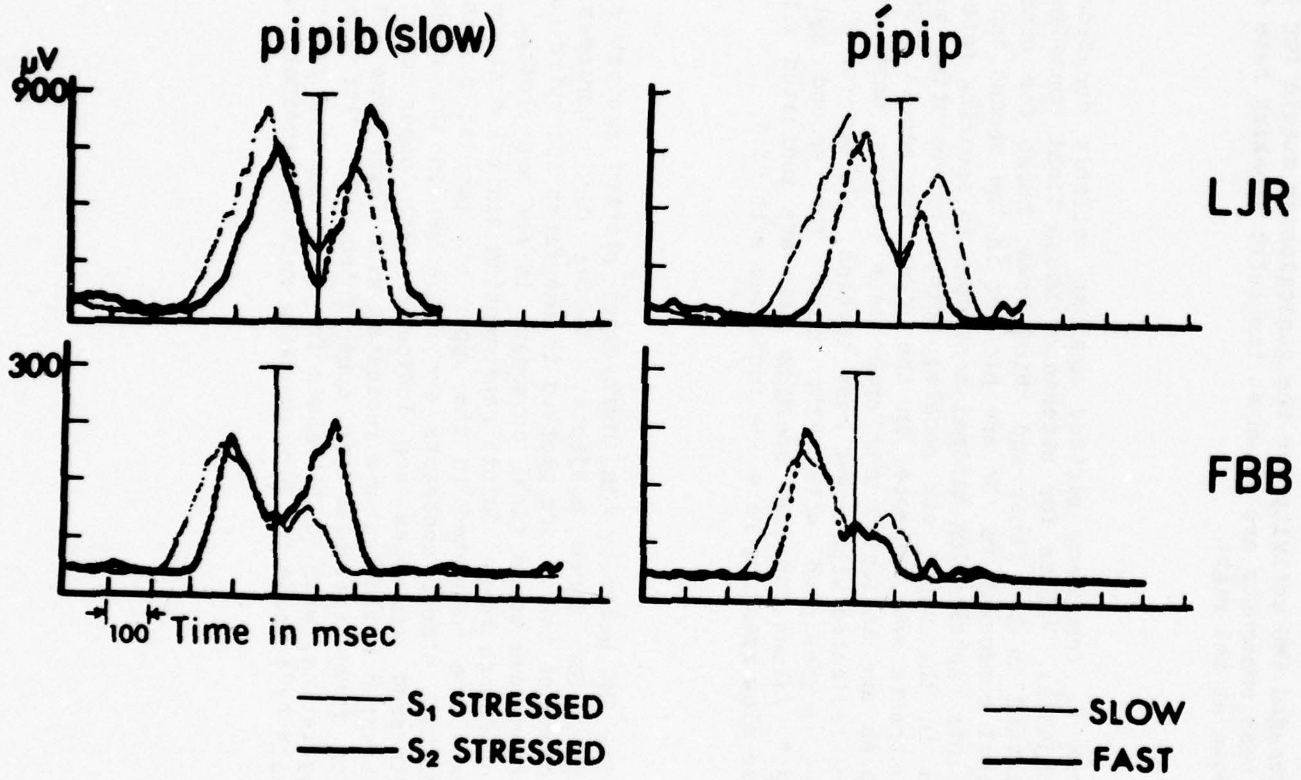

pipib(slow)
pípip
µV
900
300
LJR
FBB
100
Time in msec
S₁ STRESSED
S₂ STRESSED
SLOW
FAST

Figure 1.

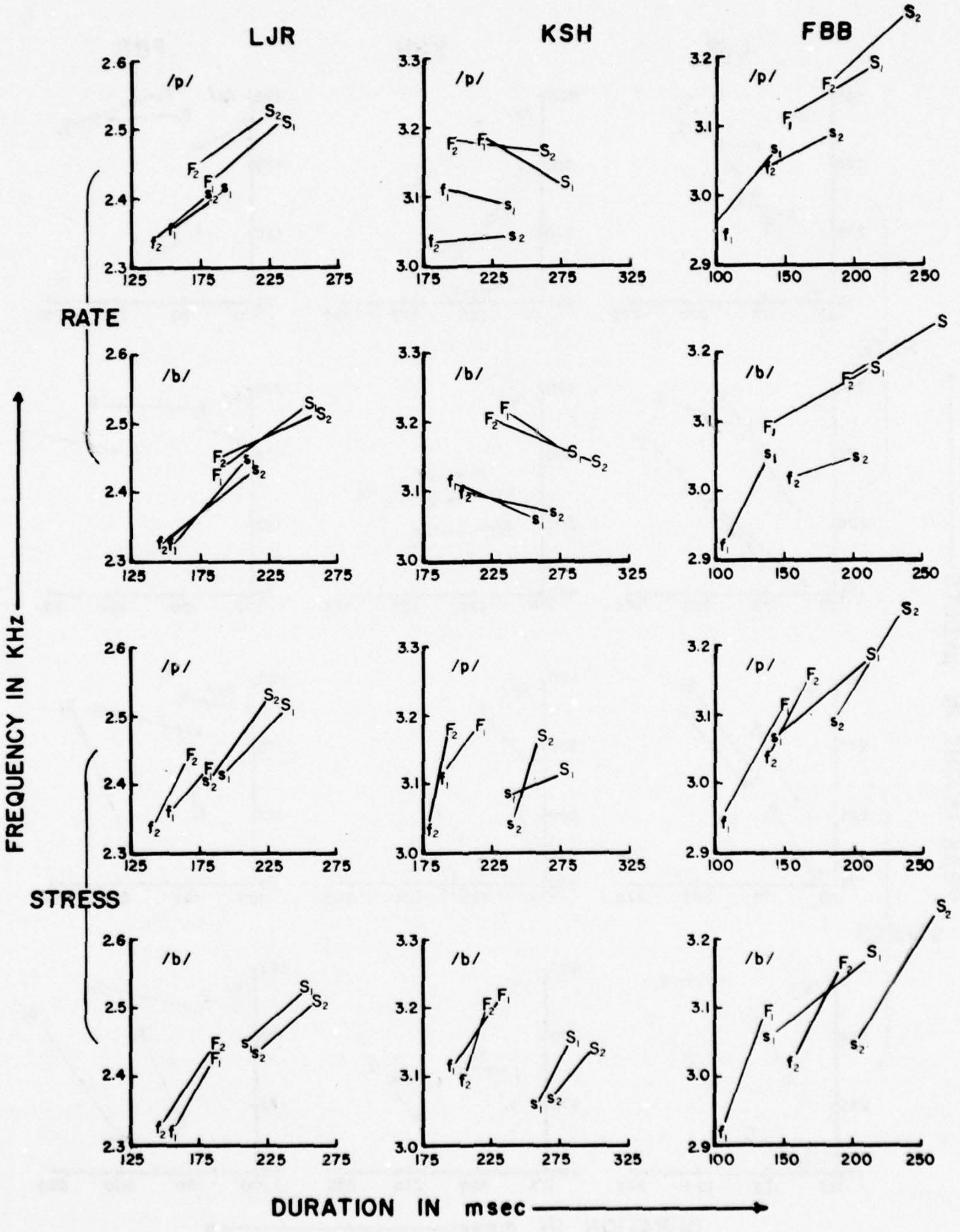

LJR
KSH
FBB
RATE
STRESS
FREQUENCY IN KHz
DURATION IN msec
/p/
/b/
2.3
2.4
2.5
2.6
3.0
3.1
3.2
3.3
2.9
125
175
225
275
325
100
150
200
250
$F_1$
$F_2$
$S_1$
$S_2$
$f_1$
$f_2$
$s_1$
$s_2$

Figure 2.

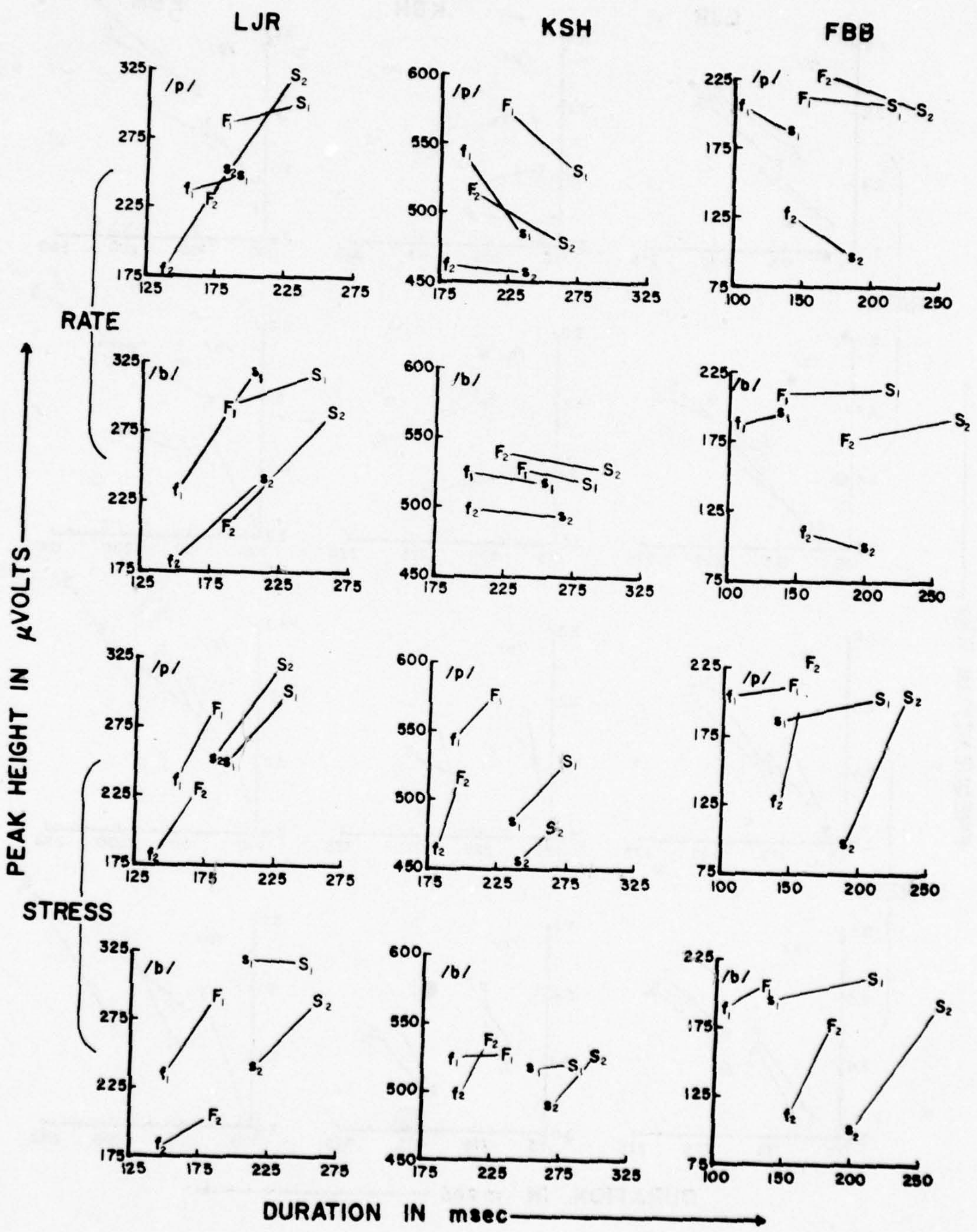

LJR
KSH
FBB
RATE
STRESS
PEAK HEIGHT IN μVOLTS
DURATION IN msec
/p/
/b/
325
275
225
175
125
600
550
500
450
325
250
200
150
100
75
F1
F2
f1
f2
S1
S2
s1
s2

Figure 3.

# THE PSYCHOLINGUISTIC BASIS OF LINGUISTIC AWARENESS[*]

Ignatius G. Mattingly[+]

__Abstract__. Two kinds of knowledge of language are distinguished: grammatical knowledge and performance knowledge. Grammatical knowledge is what a language learner acquires; performance knowledge is what a speaker-hearer uses to produce and understand sentences in real time. The reader, unlike the listener, must make direct use of both kinds of knowledge. Active language acquisition entails accessibility of grammatical knowledge. It is suggested that "linguistically aware" individuals are those who are continuing to acquire language, even though the minimal requirements of performance have been met.

If the generative grammarians (Chomsky, 1965; Chomsky & Halle, 1968) and the psycholinguists who have studied human sentence processing in the generative tradition (Fodor, Bever, & Garrett, 1974) are at least roughly correct, the speaker-hearer of a language has two distinguishable sorts of tacit knowledge that may be called grammatical knowledge and performance knowledge.

On one hand, the speaker-hearer has knowledge of the grammar of his language. The grammar consists of a lexicon and sets of ordered rules. Each entry in the lexicon is a morphophonemic representation of a word together with associated syntactic and semantic information. The lexical forms for __heal__ and __health__, for example, are /hēl/ and /hēl+θ/. The syntactic rules generate, that is, derive, the complex word-order patterns of actual sentences from elementary "deep structure" patterns by a series of structural transformations. The phonological rules generate the phonetic forms of words (which represent intended and perceived pronunciations) from morphophonemic forms by processes of substitution, insertion, and deletion. Thus the phonetic forms [hīyl] and [helθ] are derived from /hēl/ and /hēl+θ/ by rules that shorten the long vowel of /hēl+θ/ and diphthongize and shift the quality of the same vowel of /hēl/. Notice that the morphological relationship of the two words, explicit in the morphophonemic forms, becomes opaque in the phonetic forms.

Grammatical knowledge is accessible, in the sense that the speaker-hearer has intuitions about which phonetic contrasts are distinctive in his language and which are not--and which syntactic patterns are acceptable and which are

[*] This paper was presented at the Twenty-Eighth Annual Meeting of the National Reading Conference, St. Petersburg, Florida, 30 November-2 December, 1978, and has been submitted for publication in the Conference __Yearbook__.

[+] Also University of Connecticut.

__Acknowledgment__: This work was supported in part by NICHD Grant HD-01994 to Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

not. The validity of these intuitions is corroborated by the success of linguists in reconstructing descriptively adequate grammars from such intuitive data. Note, however, that there are limitations on the scope of grammatical knowledge. The speaker-hearer has very restricted intuitions, for example, about the acoustic properties of the speech signal that can be shown to determine his phonetic perceptions (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Accordingly, the grammar has nothing to say about the complex relationships between the phonetic representation of a sentence and its acoustic realization.

A child acquiring the grammar of his native language is rather in the position of a linguist. Given a theory of language, specifying the structural properties that all grammars share, and data as to correspondences between sound and meaning, he proceeds to construct the lexicon and the rules. Thus, from the phonetic forms [hīyl] and [helθ], and a notion of their meanings (as well as other parallel data), he infers the morphophonemic forms /hēl/ and /hēl+θ/ and the rules of shortening, diphthongization and vowel shift that relate the two phonological levels. (Obviously, this particular example illustrates a fairly advanced stage of language acquisition.) The child's position is different from that of the linguist mainly in that his general theory of language is innately given and superior to any general theory so far explicitly formulated by linguists. But having a task similar to the linguist's, he must have psychological mechanisms for doing what linguists do: making hypotheses about rules and about the context of lexical entries, generating hypothetical utterances and comparing them with observed utterances. Having hypothesized morphophonemic /hēl/ and /hēl+θ/ and rules of shortening, diphthongization and vowel shift, he must have a specific mechanism that enables him to generate phonetic [hīyl] and [helθ], so as to test his hypothesis.

Opposed to such grammatical knowledge about the structure of language is the performance knowledge that enables a speaker-hearer to produce and understand actual sentences in real time. What the psycholinguists have argued is that performance knowledge must be very different in form from grammatical knowledge, even though the speaker tries to produce grammatically well-formed sentences, and the listener understands sentences by constructing syntactic patterns that relate the meanings of individual words. But a speaker does not generate the sentences he utters, nor does the listener analyze the sentences he hears by applying the generative rules in reverse order. It appears, rather, the listener's sentence-processor uses various pragmatic, often fallible "parsing strategies" based on the surface orderings of possible constituents. Though there has been relatively little work on lexical search in sentence-processing, we might suppose that it is done without reference to morphophonemic forms. Phonetic forms are associated with lexical entries, and the sentence-processor, provided with a phonetic representation of an utterance by the speech-perception mechanism, finds the entries for phonetic [hīyl] and [hel+θ], and extracts the syntactic and semantic information it needs, without reconstructing morphophonemic /hēl/ and /hēl+θ/. Notice, though, that the lexicon is a link between the grammar and the sentence-processor.

It cannot be concluded, however, that grammatical knowledge is after all psychologically unreal or that is has nothing to do with understanding sentences. Which parsing strategies will be appropriate depends in part on the syntax of the language, even though the sentence-processor does not make syntactic derivations. The semantic information available to the sentence-processor in lexical entries depends in part on the semantic properties of the morphemes from which words are formed, even though it does not do morphophonemic derivations. It seems likely, therefore, that the function of grammatical knowledge is to provide the sentence-processor, in the course of language acquisition, with an optimal set of parsing strategies and an optimal set of word-meanings. Having fulfilled this function, however, grammatical knowledge has at most a vestigial role in the actual process of understanding a sentence. (See the discussion of this point in Fodor et al., 1974, pp. 368-372.)

A further difference between grammatical knowledge and performance knowledge is that although the speaker-hearer has grammatical intuitions, he does not have intuitions about performance. What is known about the procedures of the sentence-processor, therefore, has been learned by experimental inference rather than by linguistic analysis. In this respect, the sentence-processor resembles the speech perception mechanism.

How much must an actual speaker-hearer know about the grammar of his language to insure a set of parsing strategies and word-meanings sufficient for ordinary understanding? Perhaps, relatively little, in comparison with the ideal speaker-hearer of linguistic theory. It is quite believable, for example, that a person might have parsing strategies that could cope, much of the time, with passive constructions, without having grammatical knowledge of the rules for generating passive sentences;_ and that he might have a functional understanding of the meanings of [hīyl] and helθ] but no knowledge of the appropriate morphophonemic forms for these two words or of the phonological rules relating them. There is nothing in the processes of speaking and listening that compels him to learn such things.

To put the matter somewhat differently, the grammatical knowledge a language-learner is potentially capable of acquiring far exceeds the functional requirements of the sentence-processor. But if this is so, we should not find it surprising that some speaker-hearers, driven by an instinctive linguistic curiosity, continue acquiring the grammar of their language indefinitely, while others essentially abandon language acquisition once the sentence processor is adequately equipped for the purposes of ordinary communication.

Reading necessarily differs from listening because it is not possible to access performance knowledge directly with visual input. An orthography based on visual displays of acoustic wave-forms or on sound spectrograms is hardly conceivable, and even an orthography that was in effect a narrow phonetic transcription, equivalent to what a listener's speech-perception mechanism provides the sentence-processor, would be impractical. Instead, reading makes direct use of grammatical knowledge and exploits performance knowledge in a rather roundabout way.

As Chomsky (1970) points out, it is a general characteristic of orthographies that they appeal to the reader's knowledge of the morphophonemic forms of words, and not to morphologically opaque phonetic forms. In English orthography, the spellings *heal* and *health* correspond to morphophonemic /hēl/ and /hēl+θ/ and not to phonetic [hīyl] and [helθ]; and even in writing systems that leave segmental structure implicit, such as that of Chinese, it is obviously the morphological forms that are being transcribed. Apparent exceptions to this generalization--writing systems that are said to be "phonetic," turn out on examination to be used for languages that do not have highly elaborated phonologies, so that the morphophonemic forms themselves are fairly close to phonetic forms (Liberman, I. Y., Liberman, A. M., Mattingly, & Shankweiler, 1978).

The evidence of orthography suggests that a reader looks up words in his mental lexicon morphophonemically, making use of his grammatical knowledge. On the other hand, sentences are not written in a way that makes their deep structure manifest, and it does not seem likely that the reader, any more than the listener, analyzes sentences grammatically. Instead, the reader presumably uses the same parsing strategies that he uses as a listener. In order to do so, he has to provide the sentence-processor with a phonetic representation of the sentence. This can readily be done once lexical search has been accomplished, since the lexicon is linked to the sentence-processor. In the case of a known word, its phonetic form is already associated with the lexical entry, if our account of lexical search in listening is correct.

The case of an unknown word is more interesting because it provides further support for the grammatical character of reading. In this situation, the reader is obliged, having established a new lexical entry, to generate a purely hypothetical phonetic form, using the phonological rules. The mechanism that permits him to do this is not part of performance knowledge. Nor is it a special trick he learns as a reader. It is rather, we suppose, the mechanism we have already considered that enables him to test candidate morphophonemic representations in the course of language acquisition. Note that the listener, hearing a new word, is not under similar pressure to analyze it morphologically, or to use this generative mechanism.

It appears then, that the reader uses grammatical knowledge as well as performance knowledge when he reads; and that on occasion, he is obliged to rely on the same mechanisms by which this grammatical knowledge was acquired.

What has been said suggests that a good reader must know the phonology of his language to a substantially greater extent than suffices for mere speaking and listening. He is "phonologically mature" not only in the sense that the morphophonemic forms in his lexicon correspond to a large extent to the relatively abstract forms transcribed by the orthography, but also in the sense that he knows the rules that relate these forms to phonetic forms. But it would be a mistake to regard such maturity as a prerequisite for learning to read. On the contrary, as we will see, phonological maturity develops as a consequence of reading.

What does seem to be an essential prerequisite is what we have called "linguistic awareness" (Mattingly, 1972): the ability of a speaker-hearer to bring to bear rather deliberately the grammatical, and in particular, the phonological knowledge he does have in the course of reading. To the linguistically-aware child, the phonological segmentation and the morphological structure of words is intuitively obvious, and the orthography seems reasonable, even though there may be substantial discrepancies between the orthographic transcriptions of words and his immature morphophonemic forms; to the child not thus aware, the principles by which the orthography transcribes words seem quite mystifying. If the child's production and understanding of spoken language seem normal, we must suppose that he knows a reasonable amount about the phonology of his language. But he is unable to make effective use of it in reading, the task with which he is now confronted.

I believe that the considerable differences in degree of linguistic awareness that are observable in children are related to different patterns of language acquisition. As we have seen, some children seem to learn only enough of the grammar of their language to satisfy the functional requirements of the sentence-processor, and then to abandon the task of language acquisition. In such children, one might expect that at the age when reading instruction begins, grammatical knowledge would be not only more limited, but less accessible, and that acquisition mechanisms such as we have discussed would have atrophied. This mental state I would equate with lack of linguistic awareness. Such children might well have difficulty accessing the knowledge and reawakening the mechanisms required for reading, even though their speaking and listening might seem adequate.

Children of another sort, who have continued to learn the grammar of their language even after the minimal demands of the sentence-processor are satisfied, will be "linguistically aware." Their acquisition mechanisms are still in working order. Not only will they have greater phonological maturity than the first class of children, so that the orthography will correspond more closely to their mental representations of words, but also, these representations will be more accessible. Moreover, they are in a position to increase their phonological maturity. If a child does not already have the morphophonemic forms /hēl/ and /hēl+θ/ in his lexicon, and the associated rules in his phonology, he is quite likely to acquire this grammatical knowledge through reading (Moskowitz, 1973). Finally, since these children are still actively acquiring their language, they will see reading as a source of fresh data. The linguistic curiosity that motivates their continuing language acquisition will thus motivate them in learning to read as well.

My contention, then, is that linguistic awareness, essential for reading and for learning to read, is only indirectly related to speaking and understanding. It is more directly related to active language acquisition, and the awareness in question is an awareness of grammatical knowledge, of language in the form in which it is acquired. Such awareness, presumably intense in every child during the period when he learns to talk, has waned in many children by the time reading instruction begins. The reading teacher's task, essentially, is to rekindle this awareness by getting the language acquisition machinery started again.

## REFERENCES

Chomsky, N. *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press, 1965.

Chomsky, N. Phonology and reading. In H. Levin & J. P. Williams (Eds.), *Basic studies on reading*. New York: Basic Books, 1970.

Chomsky, N., & Halle, M. *The sound pattern of English*. New York: Harper and

Row, 1968.

Fodor, J. A., Bever, T. G., & Garrett, M. F. *The psychology of language*. New York: McGraw-Hill, 1974.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, I. Y., Liberman, A. M., Mattingly, I. G., & Shankweiler, D. P. Orthography and the beginning reader. *Haskins Laboratories Status Report on Speech Research*, 1979, *SR-57*.

Mattingly, I. G. Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye*. Cambridge, Mass.: MIT Press, 1972.

Moskowitz, B. A. On the status of vowel shift in English. In T. Moore (Ed.), *Cognitive development and acquisition of language*. New York: Academic Press, 1973.

# SOME EFFECTS OF LATER-OCCURRING INFORMATION ON THE PERCEPTION OF STOP CONSONANT AND SEMIVOWEL[*]

Joanne L. Miller[+] and Alvin M. Liberman[++]

Abstract: In three experiments, we determined how perception of the syllable-initial distinction between the stop consonant [b] and the semivowel [w], when cued by duration of formant transitions, is affected by parts of the sound that occur later in time. For the first experiment, we constructed four series of syllables, similar in that each had initial formant transitions ranging from one short enough for [ba] to one long enough for [wa], but different in overall syllable duration. The consequence in perception was that, as syllable duration increased, the [b-w] boundary moved toward transitions of longer duration. Then, in the second experiment, we increased the duration of the sound by adding a second syllable, [da], (thus creating [bada-wada]), and observed that lengthening the second syllable also shifted the perceived [b-w] boundary in the first syllable toward transitions of longer duration; however, this effect was small by comparison with that produced when the first syllable was lengthened equivalently. In the third experiment, we found that altering the structure of the syllable had an effect that is not to be accounted for by the concomitant change in syllable duration: lengthening the syllable by adding syllable-final transitions appropriate for the stop consonant [d] (thus creating [bad-wad]) caused the perceived [b-w] boundary to shift toward transitions of shorter duration, an effect precisely opposite to that produced when the syllable was lengthened to the same extent by adding steady-state vowel. We suggest that, in all cases, the later-occurring information specifies rate of articulation and that the effect on the earlier-occurring cue reflects an appropriate perceptual normalization.

## INTRODUCTION

In exploratory work with synthetic speech, we chanced on a phenomenon that seemed to hold promise for the study of two related perceptual effects: most directly, how later-occurring aspects of the speech signal modify the perception of an earlier-occurring cue and, by implication, how duration of

[*] Portions of the research reported in this paper were presented at the 95th meeting of the Acoustical Society of America, Providence, R.I., 1978. A slightly revised version of this manuscript is in press, Perception & Psychophysics.

[+] Northeastern University.

[++] Also Yale University and the University of Connecticut.

Acknowledgment: This research was supported by NINCDS Grant NS 14394 to J. L. Miller and NICHD Grant HD 01994 to Haskins Laboratories.

the syllable specifies, inter alia, the articulatory rate to which the listener must adjust. Following the early findings of Liberman, Delattre, Gerstman, and Cooper (1956), we had used the duration of the initial consonant-vowel transitions to produce the perceived distinction in manner between stop-vowel ([ba]) and semivowel-vowel ([wa]) syllables. Then, for purposes quite unrelated to the concerns of this paper, we varied the duration of the syllable--by extending the steady-state portion of the vowel--and observed that perception of the transition cue was, in consequence, quite markedly affected: when we made the syllable longer, the manner boundary, as we perceived it, was displaced toward a longer duration of transition.

We were reminded, then, of a series of studies by Summerfield (Note 1, Note 2) that dealt with the effect of rate of articulation on the perception of the voicing distinction in stop-vowel syllables. Having put his attention on the location of the perceptual boundary along a continuum of voice-onset-times, a major cue for the distinction in question, Summerfield found that variations in the articulatory rate of the sentence frame caused the perceptual boundary for the target phone to be displaced. More to the point of our interest here, he also found that the rate effect can be quite local, so local that it was observed when the syllable containing the target (syllable-initial) phone was isolated and the rate information was conveyed only by variations in the duration of that syllable. It is, of course, just there that Summerfield's results with voicing anticipate ours with manner.

Perhaps it is to be expected that a cue like voice-onset-time should be affected by rate of articulation, for the cue is temporal in nature. Indeed, other temporal cues have been shown to be perceived in relation to speech rate. These include, for example, the following: silence duration as a cue for voicing in intervocalic stop consonants (Port, Note 3, Note 4) and as a cue for single vs. double consonants (Pickett & Decker, 1960); frication duration and silence duration as cues for the fricative-affricate manner distinction (Dorman, Raphael, & Liberman, Note 5; Repp, Liberman, Eccardt, & Pesetsky, 1978); and vowel duration as a cue for vowel quality (Ainsworth, 1972, 1974; Verbrugge & Shankweiler, Note 6; Verbrugge, Strange, Shankweiler, & Edman, 1976). Since the transition cue for the [b-w] distinction is essentially temporal, it, too, should be perceived in relation to articulatory rate. That it is in fact so perceived is indicated by a study carried out concurrently with those we report here. That study (Minifie, Kuhl, & Stecher, Note 7) found that varying the articulatory rate of a sentence frame did affect whether a word within the sentence was heard as beginning with [b] or [w]. Moreover, investigations into the production of syllable-initial consonants at various rates of articulation have revealed changes in the speed with which the relevant gestures are made and also in the durations of the resulting acoustic transitions (Gay, 1978; Gay & Hirose, 1973; Gay, Ushijima, Hirose, & Cooper, 1974). Given that changes in transition duration occur with changes in rate, we should expect that the listener would make the appropriate normalization when using transition duration to cue a phonetic distinction.

We will here report three experiments designed to enlighten us further about the phenomenon described in our opening paragraph, namely, that the perceived distinction between syllable-initial stop and semivowel, as cued by duration of transitions, is affected by acoustic information that occurs later in time. In the first experiment, we are concerned primarily to establish the phenomenon more securely than our preliminary observations can have done. In the second experiment we ask whether the effect of later-occurring information is confined to a single syllable. The point of the third experiment is to see if the effect of the later-occurring aspects of the signal is due only to changes in syllable duration, or whether it is also influenced by the structure of the syllable.

## EXPERIMENT I

That duration of the steady-state portion of the syllable-final vowel affects the perceived boundary between syllable-initial stop and semivowel was, as we have pointed out, reasonably apparent from our initial observations. But those were based on stimulus variations that were not as systematic as they might have been, and the judgments were made only by us. It is appropriate, then, that we do the experiment properly, the more so in order to delimit the range of durations over which the effect can be found.

### Method

Stimuli. The stimuli for all experiments we report in this paper were synthetic speech patterns, generated on the Haskins Laboratories parallel-resonance synthesizer. For Experiment I, we synthesized four series of syllables that ranged from [ba] to [wa]. The syllables were three-formant patterns, consisting of a fixed initial 20 msec of prevoicing (first formant only), a variable duration of formant transition appropriate for [b] or [w], and a subsequent period of steady-state formants. To create each series, we varied the transition duration from 16 to 64 msec in four-msec steps, for a total of 13 stimuli per series. Syllables with relatively short transitions were perceived as [ba] and those with longer transitions as [wa]. As we increased the duration of transitions in four-msec steps we decreased the duration of the steady-state formants by the same amount; thus, within a given series all stimuli had the same overall duration. The four [ba-wa] series differed from each other in the overall duration of the syllables, specifically, in the duration of the steady-state formants. These syllable durations were: 80, 152, 224, and 296 msec.

For all stimuli, the first formant (F1) started at 234 Hz and rose linearly to a steady-state value of 769 Hz, while the second formant (F2) began at 616 Hz and rose linearly to a final value of 1232 Hz. The third formant (F3) remained constant at 2862 Hz. The overall amplitude of each syllable had a gradual onset, increasing by 28 dB over the course of the prevoicing and the formant transitions, and remained constant thereafter.[1]

Each of the 52 stimuli (four series X 13 stimuli per series) was digitized by the pulse code modulation (PCM) system at Haskins Laboratories. These stimuli were then used to make three randomized test orders, each containing eight instances of each stimulus. All stimuli were recorded with an interstimulus interval of three seconds.

Procedure. All subjects listened to the three test orders over the course of two sessions. They were informed that they would hear computer-generated syllables, [ba]-like or [wa]-like, of variable duration. They were asked to decide whether each syllable was [ba] or [wa], guessing if necessary,

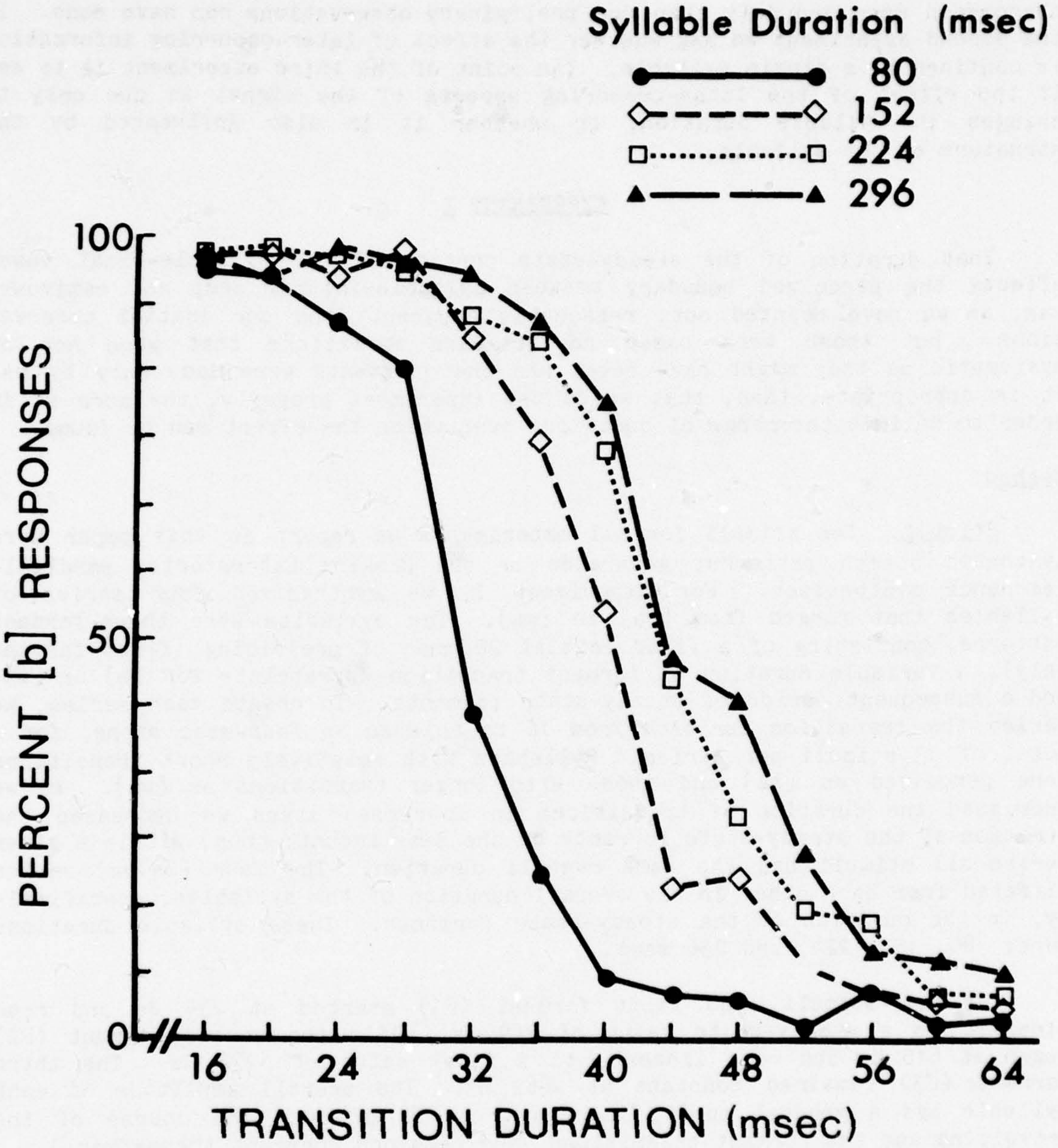


Figure 1: Effect of syllable duration on the [b-w] distinction as cued by transition duration.

and to indicate the decision by marking an appropriately formatted response sheet. The stimuli were presented to the subjects through earphones at approximately 78 dB SPL.

Subjects. The subjects were eight college students who were paid for their participation in the experiment. None reported a history of a speech or hearing disorder.

## Results

As is apparent from Figure 1, there was a systematic effect of syllable duration on the identification of the stop-consonant [b] and the semivowel [w]: as syllable duration increased, an increasingly longer transition was required to perceive [w]. To obtain a summary account of the effect, we calculated for each subject the location of the [b-w] phonetic boundary for each of the four syllable durations, using the procedure introduced by Eimas, Cooper, and Corbit (1973). That procedure calls for transforming the percentages of [b] responses to z-scores, fitting a regression line to the transformed data by finding a least-mean-squares solution, and then taking the boundary to be the stimulus value corresponding to a z-score of zero. Averaged across subjects, the mean boundary locations for the 80-, 152-, 224-, and 296-msec series proved to be at transition durations of 31.9, 41.3, 44.7, and 46.6 msec, respectively. Thus, over the range of syllable durations tested, we obtained a shift of about 15 msec in the location of the boundary. As one can see either by examining the calculated boundaries or by inspecting the identification functions of Figure 1, the largest boundary shifts were at the shortest syllable durations, with increasingly smaller shifts occurring as syllable duration increased. A consequence is that the gap between the functions for the 80- and 152-msec series--about nine msec of difference in the location of the boundary--is quite large.

In order to fill that gap, and at the same time to test the replicability of our initial findings, we conducted an auxiliary study in which we included the two shortest syllable durations (80 and 152 msec) of the initial study, with a new condition of syllable duration (116 msec) lying midway between them. (These new syllables were constructed by extending the steady-state vowel of each syllable in the 80-msec series by 36 msec.) The subjects for this auxiliary experiment were the same eight students who had participated in the initial one. In Figure 2 we see the results of the three syllable-duration conditions of the auxiliary experiment and, for comparison, the results of the two corresponding conditions of the initial study. It is plain that the two corresponding conditions of the two studies did indeed produce essentially the same results and that the new, intermediate condition produced an intermediate result.

The combined data from the two studies, shown in Figure 3, clearly indicate that, as syllable duration increases, there is a perfectly regular change in the way our subjects identified the patterns as [ba] and [wa]: the longer the duration of the syllable, the longer the duration of transition needed in order to hear [wa]. Using the same combined data, we calculated, for each subject, the location of the [b-w] boundary for each of the five syllable durations. Those data are presented in Table 1, where we see that the effect of syllable duration, shown in Figure 3, occurred for every subject

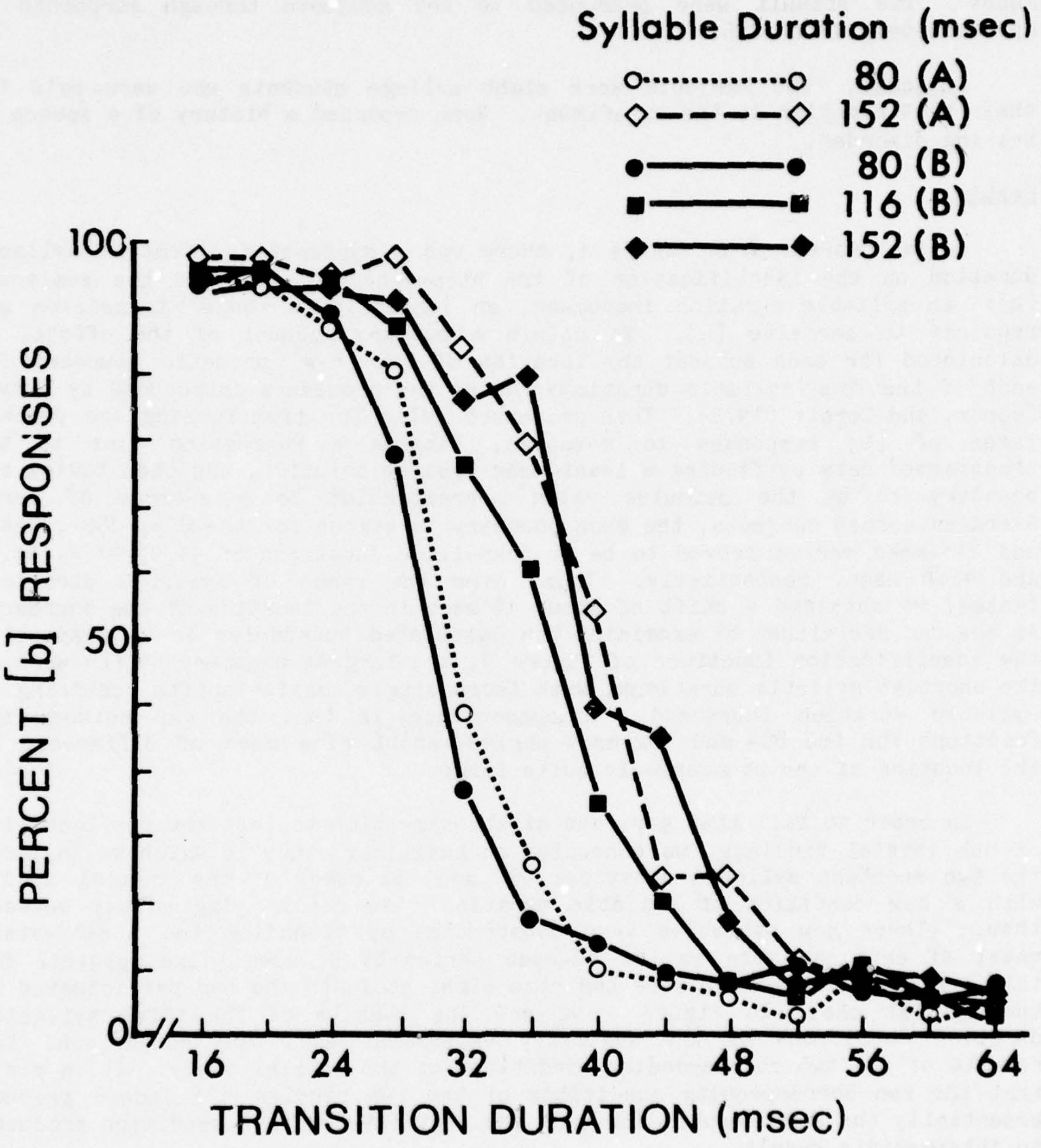


Figure 2: Replication and extension of the effect of syllable duration on the [b-w] distinction. The data from the main study are shown by open symbols and the data from the auxiliary study by filled symbols.

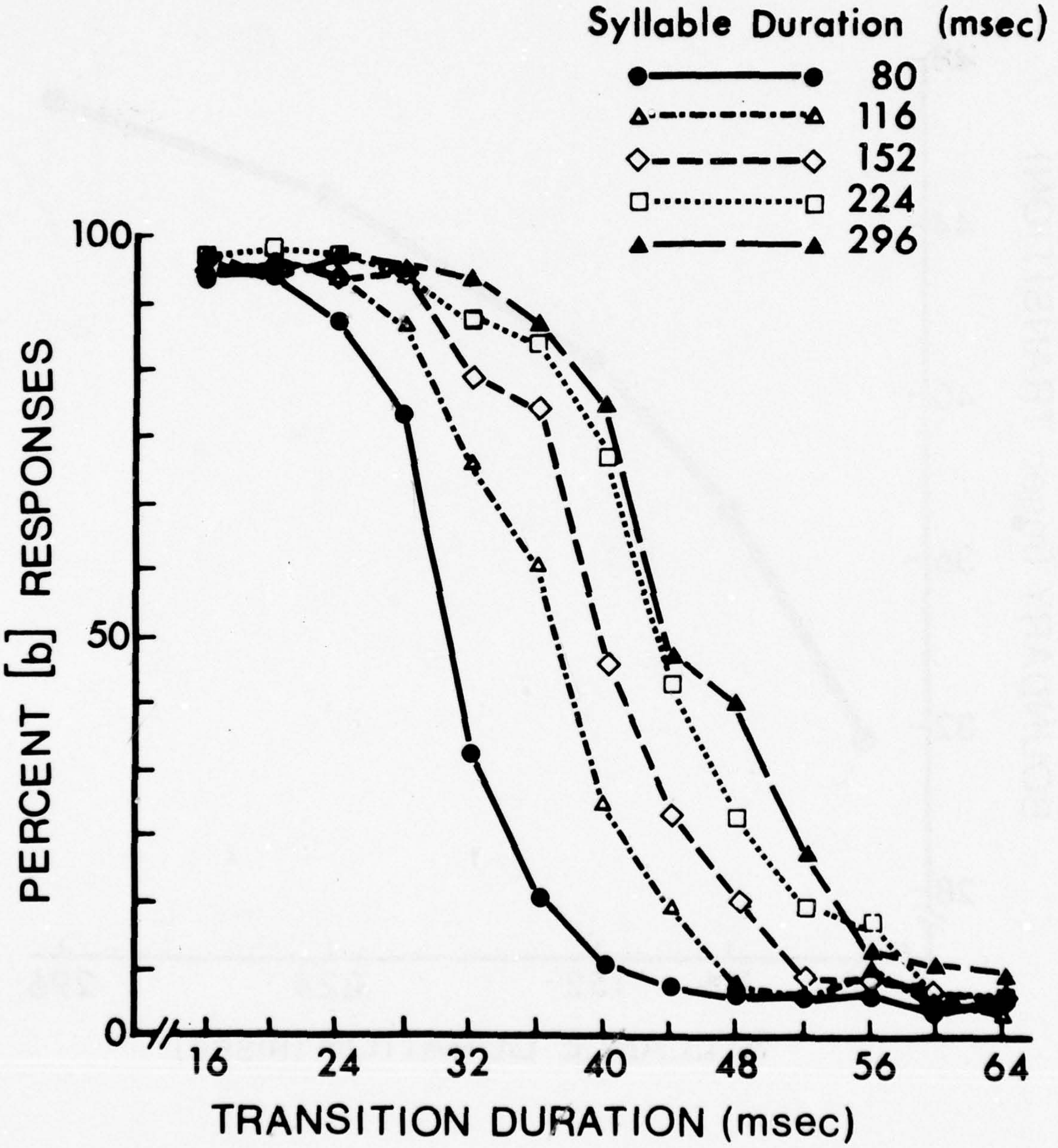


Figure 3: Combined results of the two studies of the effect of syllable duration on the [b-w] distinction. (The identification functions *for the* 80- and 152-msec series are based on data from both studies of Experiment I.)

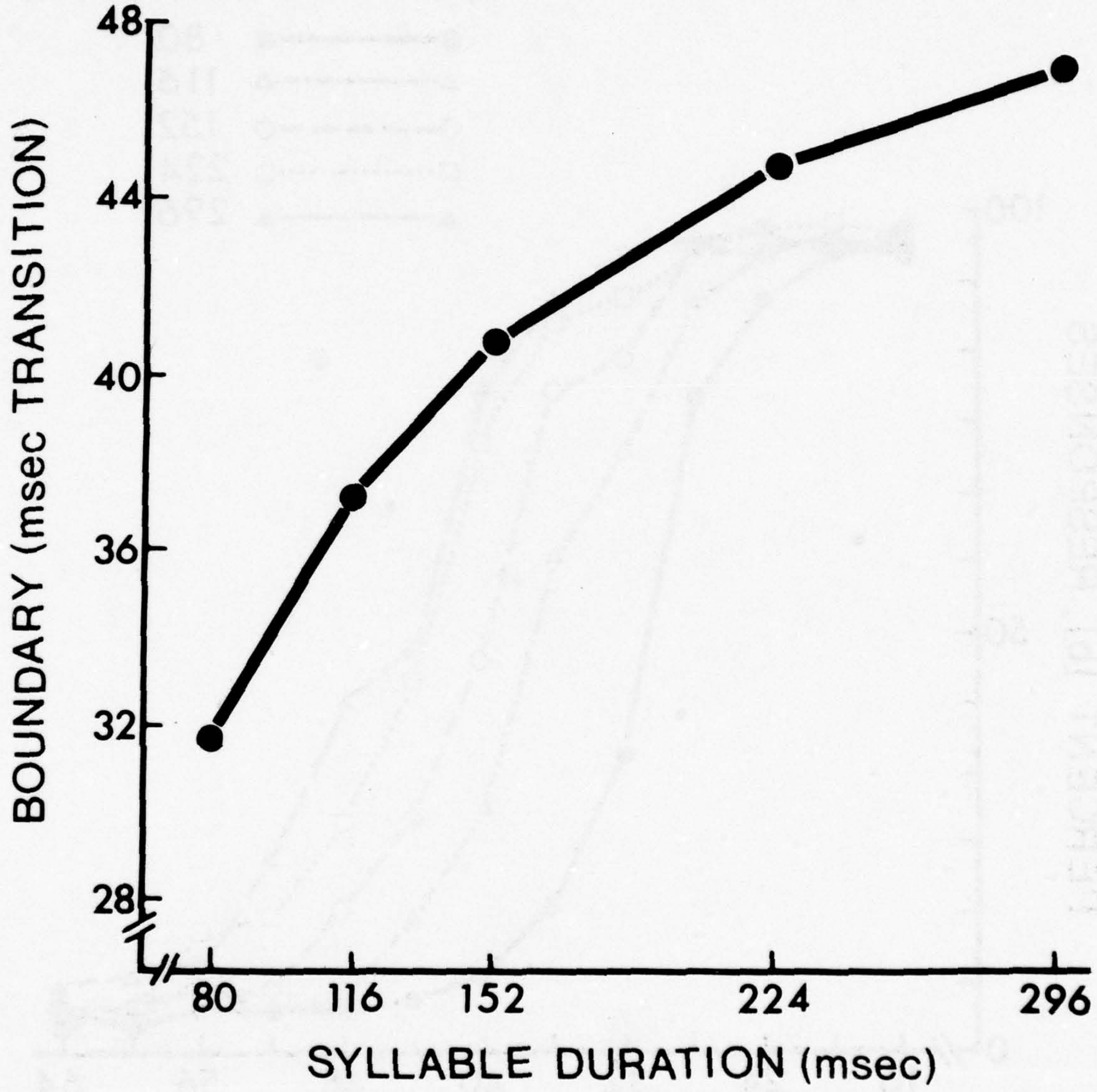


Figure 4: Effect of syllable duration on the location of the [b-w] boundary. (The boundary values for the 80- and 152-msec series are based on data from both studies of Experiment I.)

in nearly every syllable duration condition. The mean boundary values are shown, as a function of syllable duration, in Figure 4. There we see, as we might have inferred from Figure 3, that the transition duration at the boundary is a smooth and negatively accelerated function of the duration of the syllable.

TABLE 1: Individual and mean [b - w] boundary values, in msec of transition duration, for the several syllable durations of Experiment I. (The scores for the 80 and 152 msec durations are the average of the scores from the main and auxiliary studies.)

| | Syllable Duration | | | | |
|---|---|---|---|---|---|
| Subject | 80 | 116 | 152 | 224 | 296 |
| 1 | 28.8 | 35.2 | 40.0 | 47.2 | 48.0 |
| 2 | 33.6 | 36.0 | 40.8 | 43.2 | 46.4 |
| 3 | 38.4 | 45.6 | 46.4 | 48.8 | 48.0 |
| 4 | 28.8 | 32.0 | 37.6 | 40.8 | 41.6 |
| 5 | 28.8 | 36.0 | 39.2 | 45.6 | 46.4 |
| 6 | 33.6 | 36.0 | 42.4 | 48.8 | 50.4 |
| 7 | 28.8 | 35.2 | 39.2 | 44.0 | 44.8 |
| 8 | 33.6 | 41.6 | 41.6 | 39.2 | 47.2 |
| $\bar{x}$ | 31.8 | 37.2 | 40.9 | 44.7 | 46.6 |

## EXPERIMENT II

In Experiment I we found that later-occurring information in the syllable affected the perception of an earlier-occurring cue: the duration of the syllable determined whether the initial *formant* transitions of different durations were perceived as [b] or [w]. Now we mean to find out whether such effects are contained within syllable boundaries. What is the effect, if any, of adding a second syllable of variable duration to the one containing the transition cue?

There is reason to believe that the effect of adding a second syllable will be small. We have in mind the experiments by Summerfield (Note 1, Note 2) described in the introduction to this paper. As the reader may recall, Summerfield found that perception of an important temporal cue for voicing (voice-onset-time) was affected by variations in the rate at which the utterance was articulated and, further, that the effect was quite local. That is, rate information closer to the target phone had more effect on its perception than rate information that occurred farther away. A similar result has been reported by Port (Note 4) for the intervocalic voicing distinction as cued by duration of intersyllabic silence. If, concerning the results of Experiment I, we assume that syllable duration had its effect because it specified the rate of articulation, then we might suppose that, following the Summerfield and Port studies, the most important duration would be that of the syllable containing the target phone.

## Method

Stimuli. For this experiment we used stimuli that were identical to some of those used in Experiment I, except that a [da] was added to each syllable. To create these disyllables, we selected from the stimuli of Experiment I just those that had durations of 80 and 224 msec. The reader will recall that stimuli of both durations--let us call them the 80-msec series and the 224-msec series--had initial formant transitions that ranged from 16 to 64 msec in steps of four msec, and were perceived, depending on the duration of the formant transitions, as [ba] or [wa]. For the purposes of this experiment, we added to each of those patterns a synthetic syllable, [da], 72 msec in duration. Thus, we had one series containing stimuli composed of an 80-msec [ba] or [wa] followed by a 72-msec [da], which we will refer to as the 80-72 msec series, and one series with stimuli consisting of a 224-msec [ba] or [wa] followed by a 72-msec [da], which we will refer to as the 224-72 msec series. Next, we created two additional [bada-wada] series by extending the steady-state formants of the [da] so as to make the syllable 216 (instead of 72) msec long. These will be referred to as the 80-216 msec series and the 224-216 msec series. Note that a comparison of performance on the two series containing a short [ba] or [wa] (the 80-72 and 80-216 msec series) with performance on the two series containing a long [ba] or [wa] (the 224-72 and 224-216 msec series) will allow us to assess the effect of lengthening the syllable containing the transition cue for [b-w]. On the other hand, comparing performance on the two series containing the short [da] (the 80-72 and 224-72 msec series) with that on the series containing the long [da] (the 80-216 and 224-216 msec series) will show the effect of lengthening not the target syllable, but the one following.

Each of the two [da]'s contained an initial 24 msec of transition followed by steady-state formants, 48 msec in length for the 72-msec [da] and 192 msec in length for the 216-msec [da]. The starting frequency values for the [da] transitions were 234 Hz (F1), 1541 Hz (F2), and 3195 Hz (F3), and the steady-state formant frequency values for the [da] were the same as those for the [ba] or [wa], namely, 769 Hz (F1), 1232 Hz (F2), and 2862 Hz (F3). As for the amplitude of the [da], it increased by 10.5 dB over the initial transition segment, reaching and maintaining a level equal to that of the first syllable ([ba]) or ([wa]). Fundamental frequency for [da] was also set equal to that of the first syllable.

Each of the 52 stimuli (four series X 13 stimuli per series) was digitized using the PCM system. We then generated three randomized test orders, each containing eight instances of each of these 52 tokens. These orders were recorded on audio tape with an interstimulus interval of three seconds.

Procedure. The subjects were presented with the three test orders over the course of two sessions. They were informed that they would hear computer-generated disyllables, [bada] or [wada], and that the durations of both syllables would vary. They were asked to decide whether the first syllable of each stimulus was [ba] or [wa], guessing if necesssary, and to indicate their choice by writing B or W on an answer sheet. All subjects heard the stimuli through earphones at approximately 78 dB SPL.

Subjects. Fourteen paid listeners participated in this experiment, including three subjects who served as listeners in the first experiment. All were college students or staff who reported no history of a speech or hearing disorder.

## Results

In Figure 5 are plotted the data from the four [bada-wada] series. We should first examine the effect on the [b-w] boundary of changing the duration of the first syllable--that is, the one containing the target [b] or [w] phone--while holding constant the duration of the second syllable ([da]). Clearly, lengthening the first syllable (from 80 to 224 msec) shifted the [ba-wa] boundary to a longer duration of transition, and this was true whether the duration of the second syllable was 72 or 216 msec. Calculating the boundary values by the method described in Experiment I, we obtained values (in msec of transition) for the four disyllables as follows: 80-72 = 33.0, 80-216 = 35.9, 224-72 = 40.7, and 224-216 = 41.9. The average boundary value for the two series with a short [ba] or [wa] (80-72 and 80-216) is 34.4 msec, and that for the two series with a long [ba] or [wa] (224-72 and 224-216) is 41.3 msec. Thus the boundary shift attributable to the difference in duration of the first syllable is about seven msec. The reliability of this difference was confirmed by an analysis of variance, First Syllable X Second Syllable X Subject, performed on the boundary scores, that showed a significant effect of First Syllable ($p < .001$) but no significant interaction. Thus, the magnitude of shift resulting from lengthening the first syllable did not depend on the duration of the second syllable.

Consider next the effect on the [b-w] boundary of changing the duration of the second syllable ([da]) from 72 to 216 msec. Making the appropriate comparisons in Figure 5, we see that there was, indeed, an effect and, further, that the effect was in the same direction as that produced by variation in the duration of the first syllable. Moreover, the analysis of variance on the boundary scores, mentioned above, showed that the effect of lengthening the second syllable was reliable ($p < .01$). Again, the lack of a reliable interaction between First Syllable and Second Syllable indicated that the effect of lengthening [da] did not differ as a function of the duration of the first syllable. Of particular relevance to our purposes, however, is the fact that the boundary shift produced by changing the duration of the second syllable, when averaged over the two durations of the first syllable, proved to be only two msec, smaller by a fair margin than the seven-msec boundary shift produced by varying the duration of the first syllable. A difference in that direction was in fact found with 13 of the 14 subjects; the remaining subject showed no difference.

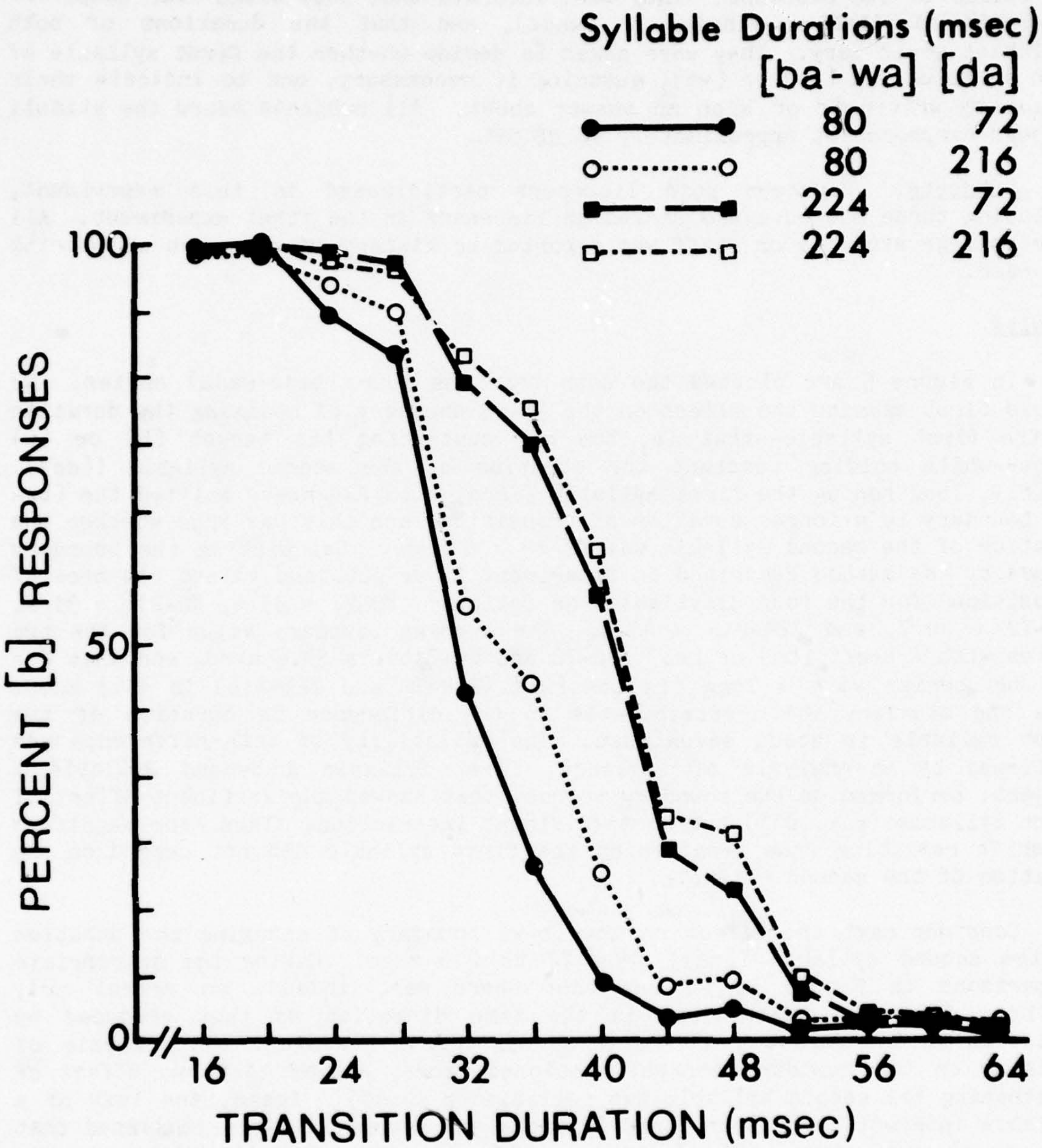


Figure 5: Effect of duration of a first and second syllable on perception of the [b-w] distinction in initial position in the first syllable.

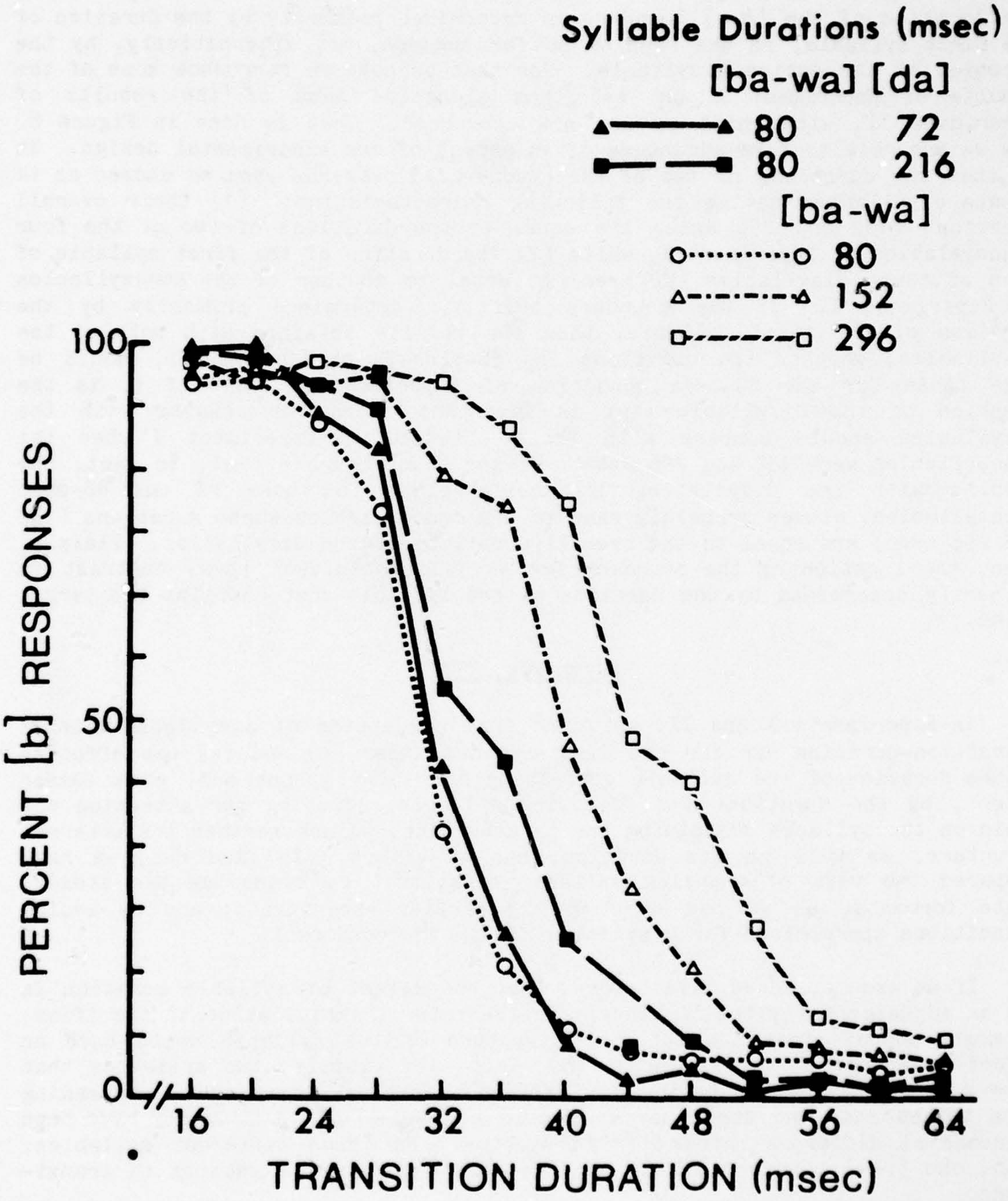


Figure 6: Comparison of the effect of varying the duration of a first and second syllable on perception of the [b-w] distinction in initial position in the first syllable.

We move now to comparisons that permit us to see more directly whether the location of the [b-w] boundary is determined primarily by the duration of the first syllable, as the results so far suggest, or, alternatively, by the duration of the entire disyllable. For that purpose we reproduce some of the results of Experiment I and set them alongside some of the results of Experiment II, with which we are now concerned. That is done in Figure 6. Now we are able to take advantage of an aspect of our experimental design. It is that the durations of two of the [bada-wada] patterns were so chosen as to create disyllables having the following characteristics: (1) their overall durations (152 and 296 msec) are equal to the durations of two of the four monosyllables of Experiment I, while (2) the duration of the first syllable of each of these disyllables (80 msec) is equal to another of the monosyllables of Experiment I. If the boundary shift is determined primarily by the duration of the first syllable, then the results obtained with both of the disyllables, whether the durations are 80-plus-72 or 80-plus-216, would be like those for the 80-msec condition of Experiment I. But if it is the duration of the disyllable that is important, then the results with the disyllables should compare with those obtained in Experiment I when the monosyllables were 152 and 296 msec. We see from Figure 6 that, in fact, the results with the disyllables lie quite close to those of the 80-msec monosyllables, closer certainly than to the monosyllables whose durations (152 and 296 msec) are equal to the overall durations of the disyllables. Plainly, then, the location of the boundary for a syllable-initial [b-w] contrast is primarily determined by the duration of the syllable that contains the target phone.

## EXPERIMENT III

In Experiments I and II, we found that perception of a syllable-initial transition-duration cue for the distinction between [b] and [w] was affected by the duration of the syllable containing the target phone and, to a lesser extent, by the duration of a following syllable. Putting our attention now again on the syllable containing the target phone, we ask whether its internal structure, as well as its duration, has an effect. To find out, we have compared two ways of changing syllable duration: by extending the steady-state formants, as we had done in the earlier experiments, and by adding transitions appropriate for a syllable-final stop consonant.

If we assume, as we have before, that the effect of syllable duration is via an adjustment (by the listener) for the rate of articulation it specifies, we might suppose that the internal structure of the syllable would have an effect independently of its duration. Thus, for example, two syllables that have the same overall duration but different internal structures--one ending in a voiced stop and the other ending in a vowel--would presumably have been produced at different rates of articulation. For such different syllables, then, the [b-w] boundary should be located at different durations of transitions.

### Method

Stimuli. The stimuli for this study comprised four series of synthetic syllables. Two of these, identical with some that were used in Experiment I, consisted of prevoicing and initial formant transitions (appropriate for [b] or [w]) of variable duration (16 msec to 64 msec in steps of four msec), followed by steady-state formants (appropriate for [a]) of variable duration, to yield CV syllables with total stimulus durations of 80 msec in the case of one series and 116 msec in the other. We will refer to these "old" series as the [ba-wa]-80 and the [ba-wa]-116 series. The two new series were formed by simply adding to the end of each of the "old" patterns 36 msec of formant transitions appropriate for a syllable-final [d]. Consequently, these syllables sounded like [bad] or [wad] and had a duration of 116 msec in the case of the one series ([bad-wad]-116) and 152 msec in the other ([bad-wad]-152). Across the final [d] transition, the first formant fell linearly from its steady-state value of 769 Hz to 234 Hz, as the second and third formants rose linearly from their steady-state levels of 1232 Hz and 2862 Hz to 1541 Hz and 3363 Hz, respectively. During this period, the overall amplitude fell 4.5 dB.

Using the PCM system, we created three test orders, each containing random arrangements of eight tokens of each of the 52 stimulus types (four series X 13 stimuli per series). The interstimulus interval was three seconds.

Procedure. The three test orders were presented to the subjects in two sessions. Subjects were told that they would hear one of four syllables--[ba], [wa], [bad], or [wad]--and that these would vary in duration. They were asked to indicate for each syllable whether it began with [b] or [w], and to guess if necessary. The syllables were presented through earphones at approximately 78 dB SPL.

Subjects. The subjects were ten paid college students who reported no speech or hearing disorders. Two of the listeners had participated in one or more of our previous experiments with [b-w].

## Results

Having in mind that half of this experiment was an exact repetition of Experiment I, we should first make the appropriate comparison of results. For that purpose we look, in Figure 7, at the functions that were obtained with the two [ba-wa] series. We observe that, as in Experiment I, lengthening the steady-state vowel caused the [b-w] boundary to shift toward a longer duration of transition. As determined by the method referred to in Experiment I, the phonetic boundaries were found to lie at 36.1 msec for the syllables with overall durations of 80 msec ([ba-wa]-80) and at 41.8 msec for those with durations of 116 msec ([ba-wa]-116). These are to be compared with boundaries of 31.8 msec and 37.2 msec that were obtained for the same conditions in Experiment I (see Table 1). Thus, the magnitude of the boundary shift owing to the 36 msec change in syllable duration was approximately equal in the two experiments, but in Experiment III all boundaries fell at longer transition durations. We do not know why. The only differences were in the subjects and in the overall contexts in which the stimuli were presented.

In all of the experimental results so far presented, the target syllable--that is, the one containing the syllable-initial [b] or [w]--had the structure CV. It is of interest, then, to examine the effect of varying the duration of the steady-state vowel in a CVC syllable. To see that effect, we look at the two remaining functions in Figure 7--that is, those for the [bad-wad] series

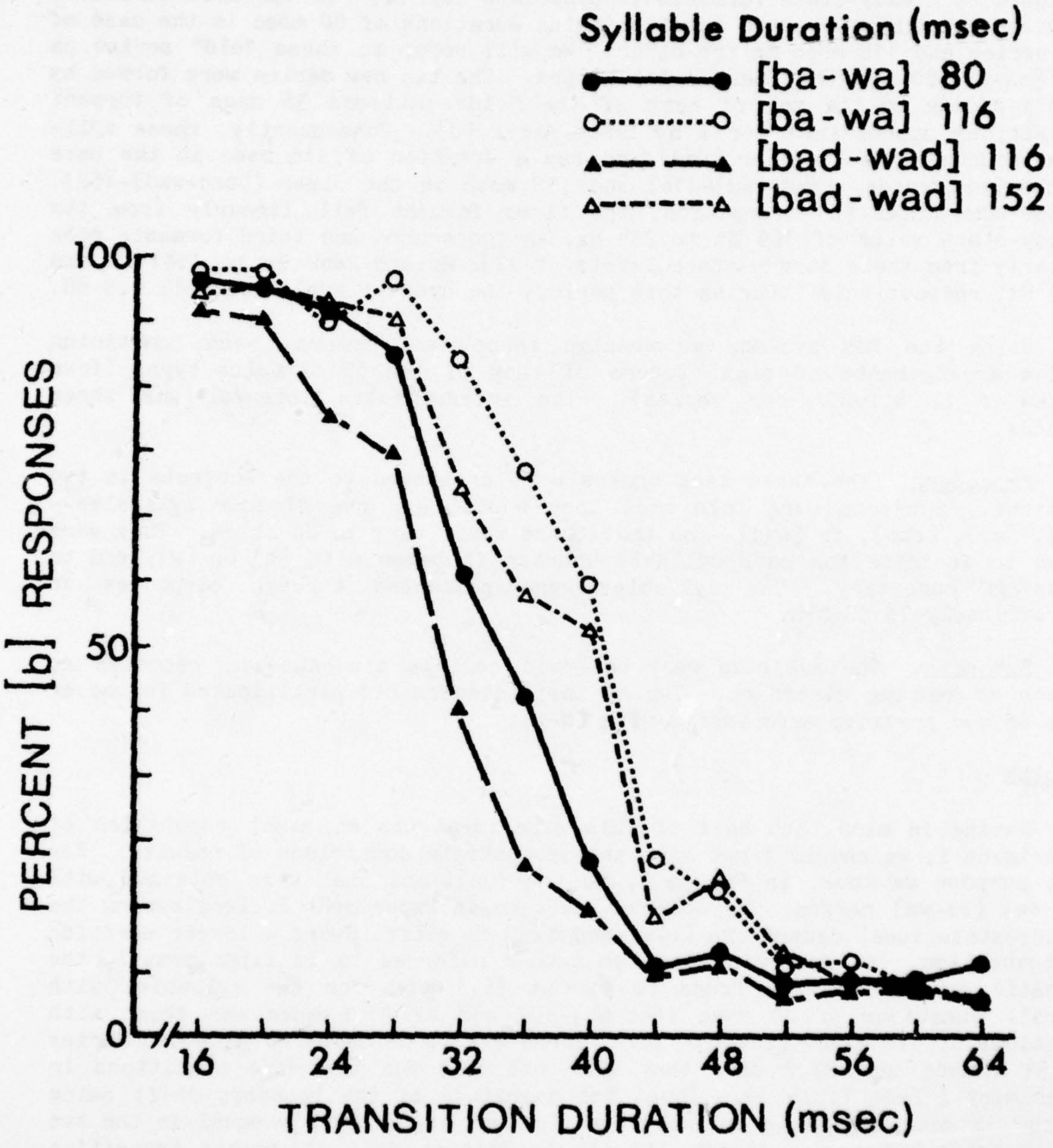


Figure 7: Comparison of the effect on the [b-w] distinction of lengthening the syllable by extending the steady-state formants and by adding transitions that cue a final consonant.

having durations of 116 msec ([bad-wad]-116) and 152 msec ([bad-wad]-152). We observe that with the CVC syllable, as with the CV, increasing the duration of the steady-state vowel causes the [b-w] boundary to shift toward longer durations of transition. Calculating these boundaries as we have the others, we obtained locations of 32.5 msec and 39.4 msec for the [bad-wad]-116 and [bad-wad]-152 series, respectively.

Having now seen the effect of increasing the duration of the syllable by adding steady-state vowel (in both CV and CVC structures), we can turn to the question that is of greatest interest to us in this experiment--namely, how does this effect compare with that which is obtained when the same increases in syllable duration are produced by adding, not steady-state vowel, but syllable-final formant transitions appropriate for a stop? To answer that question, we should use the functions shown in Figure 7 to make two comparisons. The first is between [ba-wa]-80 and [bad-wad]-116. We see, then, that the phonetic boundary for [bad-wad]-116 is at a _shorter_ transition duration than that for [ba-wa]-80. Recall, now, that the effect of increasing syllable duration by adding 36 msec of steady-state vowel was to shift the phonetic boundary toward a _longer_ duration of syllable-initial transition. Thus, the two ways of increasing syllable-duration--adding steady-state vowel and adding syllable-final (stop) transitions--have exactly opposite effects.

The second comparison we want to make is between [ba-wa]-116 and [bad-wad]-152. Here we see the same effect that we observed in the comparison we just made between [ba-wa]-80 and [bad-wad]-116--namely, that lengthening the syllable by adding 36 msec of syllable-final transitions shifted the phonetic boundary for the [b-w] distinction toward shorter durations of syllable-initial transitions. Since we do not have in this experiment a condition of [ba-wa]-152, we cannot directly compare the effect of lengthening the syllable by the two different means, as we did above for the cases of [ba-wa]-80, [ba-wa]-116, and [bad-wad]-116. However, we know from the results of Experiment I that changing syllable duration from 116 to 152 msec by adding 36 msec of steady-state vowel did, in fact, shift the boundary toward longer durations of transitions. Thus, we have further evidence that adding 36 msec of steady-state vowel and the same amount of syllable-final transition have opposite effects.[2] This conclusion is supported by an analysis of variance performed on the individual boundary scores, Adding Steady-State Vowel X Adding Formant Transitions X Subject. The opposite effects of Adding Steady-State Vowel and Adding Formant Transitions were both significant ($p < .001$ and $p < .05$, respectively), and the interaction between these two effects was not significant ($p > .10$).

## DISCUSSION

Our three experiments show that information occurring later in the speech stream affects the perception of an earlier-occurring cue. When considered, most generally, as an after-going effect, our finding is one of a class that is common in speech perception. For example, it has been found by several investigators that perception of cues for stop and fricative consonants depends in some instances on the nature of the following vowel (e.g., Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Dorman, Studdert-Kennedy, & Raphael, 1977; Fischer-Jørgensen, 1954; Kunisaki & Fujisaki, 1977; Repp & Mann, Note 8). Moreover, the effect is not confined within a single syllable, since information present in a following syllable can affect the perception of phonetic segments that belong to the preceding one (e.g., Dorman, Raphael, Liberman, & Repp, Note 9; Repp, 1978; Repp et al., 1978). Such after-going effects are important, because they imply that phonetic perception is not accomplished, phone by phone, in a simple progression through the acoustic signal. Apparently, the perceiver operates over relatively long stretches of sound, integrating into unitary phonetic percepts a numerous variety of acoustic cues that are quite widely distributed in time and thoroughly overlapped with cues for other phones (cf. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

Of the after-going effects reported in the literature, most are associated with the way articulatory and coarticulatory maneuvers smear the acoustic information for the discrete and successive segments of the phonetic message. Thus, the cues for a single phone may be spread through several acoustic segments as, for example, when stop-consonant closure and opening into a following vowel produce a period of silence, a transient burst of sound, a period of frication, some aspiration, and, finally, the onset of voicing, usually at some point during the formant transitions into the vowel (cf. Fant, 1973; Fischer-Jørgensen, 1954; Halle, Hughes, & Radley, 1957). Or, in apparently opposite fashion, coarticulation may cause cues for successive phonetic segments to be collapsed into a single segment of sound and conveyed simultaneously on the same acoustic parameter as, for example, when the initial formant transitions convey information both about the initial consonant and the following vowel of a CV syllable (Liberman et al., 1967).

The after-going effects we have reported in this paper are, however, of a different sort, in that they are apparently owing to a different cause: normalization for the consequences of changes in rate of articulation. As we pointed out in the introduction to this paper, there is reason to believe that different rates of articulation produce different durations of formant transition in syllable-initial consonants. There is, moreover, considerable basis for supposing that information about rate of articulation is provided by syllable duration, one of the variables of our experiment (cf. Gaitenby, Note 10; Gay, 1978; Klatt, 1976; Peterson & Lehiste, 1960). We conclude, therefore, that the result of the first two experiments--namely, that the perceived [b-w] boundary shifted as a function of syllable-duration--is to be interpreted as an appropriate adjustment by the listener for changes in articulatory rate.

Having in mind that one of the rate-specifying variables in our experiments was simply duration of the syllable, we should take note here of those cases in which duration is a cue in its own right. There is, for example, the distinction between voiced and voiceless stops in syllable-final position (e.g., [ɛd] vs. [ɛt]). In that case, the syllable is longer, other things equal, when the final stop is voiced, and there is evidence that listeners use the duration appropriately in identifying the voicing value of the final segment (e.g., Denes, 1955; Raphael, 1972). Or, as is well known, a similar situation exists for certain vowel distinctions. Thus, [æ] is inherently longer than [ɛ] and here, too, duration is, per se, a cue for the phonetic identity of the vowel (e.g., Mermelstein, Liberman, & Fowler, Note 11; Peterson & Lehiste, 1960; Verbrugge & Shankweiler, Note 6).

Thus duration can, in fact, specify duration and not, as in our experiment, rate of articulation. Accordingly, it is of interest to know what happens in such cases when, as in our third experiment, duration is increased not by extending the steady-state vowel but by adding formant transitions appropriate for a stop consonant. Such experiments have shown that adding (stop consonant) formant transitions has the same effect as adding a certain duration of steady-state vowel (Mermelstein et al., Note 11; Raphael, Dorman, & Liberman, Note 12). That is, when duration, qua duration, is being specified, formant transitions contribute to it, just as we should expect, given the impossibility of dividing a syllable into acoustic segments (transitions and steady-state) that correspond one-to-one with the phonetic segments. In contrast, as the reader will recall, we found in our third study that increasing the duration of the syllable by adding formant transitions had an effect precisely opposite to that produced by adding the same amount of steady-state.[3] This suggests that, in our experiments, duration itself was not the cue; its effect was presumably owing instead to its role in specifying rate of articulation. In that case, we should have expected that the structure of the syllable as well as its duration would be important.

The experiments reported here have demonstrated that the effect of transition duration as a cue for the [b-w] distinction is influenced by the duration and structure of the syllable containing the cue and, to a lesser extent, by the duration of a subsequent syllable. In our view, this aftergoing effect reflects an adjustment by the listener to the articulatory rate of the speaker: the duration and structure of the syllable provide information about rate, and the listener uses this information when making a phonetic judgment of [b] vs. [w].

## REFERENCE NOTES

1. Summerfield, A. Q. Information processing analyses of perceptual adjustments to source and context variables in speech. Unpublished Ph.D. thesis, The Queen's University of Belfast, 1975.
2. Summerfield, A. Q. On articulatory rate and perceptual constancy in phonetic perception. Unpublished manuscript, 1978.
3. Port, R. F. The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Unpublished Ph.D. thesis, University of Connecticut, 1976.
4. Port, R. F. Effects of word-internal versus word-external tempo on the voicing boundary for medial-stop closure. Paper presented at the 95th meeting of the Acoustical Society of America, Providence, 1978.
5. Dorman, M. F., Raphael, L. J., & Liberman, A. M. Further observations on the role of silence in the perception of stop consonants. Paper presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., 1976.
6. Verbrugge, R. R., & Shankweiler, D. P. Prosodic information for vowel identity. Paper presented at the 93rd meeting of the Acoustical Society of America, University Park, 1977.
7. Minifie, F., Kuhl, P., & Stecher, B. Categorical perception of [b] and [w] during changes in rate of utterance. Paper presented at the 94th meeting of the Acoustical Society of America, Miami, 1977.
8. Repp, B. H., & Mann, V. A. Influence of vocalic context on perception of the [s] - [ʃ] distinction. Paper presented at the 96th meeting of the

Acoustical Society of America, Honolulu, 1978.

9. Dorman, M. F., Raphael, L. J., Liberman, A. M., & Repp, B. Some masking-like phenomena in speech perception. Paper presented at the 89th meeting of the Acoustical Society of America, Austin, 1975.

10. Gaitenby, J. The elastic word. In *Haskins Laboratories Status Report on Speech Research*, 1965, *SR-2*, 3.1-3.12.

11. Mermelstein, P., Liberman, A. M., & Fowler, A. Perceptual assessment of vowel duration in consonantal context and its application to vowel identification. Paper presented at the 94th meeting of the Acoustical Society of America, Miami, 1977.

12. Raphael, L. J., Dorman, M. F., & Liberman, A. M. The perception of vowel duration in VC and CVC syllables. Paper presented at the 89th meeting of the Acoustical Society of America, Austin, 1975.

## REFERENCES

Ainsworth, W. A. Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*, 1972, *51*, 648-651.

Ainsworth, W. A. The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, 1974, *17*, 103-109.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, 1952, *24*, 597-606.

Denes, P. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 1955, *27*, 761-764.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 1977, *22*, 109-122.

Eimas, P. D., Cooper, W. E., & Corbit, J. D. Some properties of linguistic feature detectors. *Perception & Psychophysics*, 1973, *13*, 247-252.

Fant, G. *Speech sounds and features*. Cambridge: MIT Press, 1973.

Fischer-Jørgensen, E. Acoustic analysis of stop consonants. *Miscellanea Phonetica*, 1954, *2*, 42-59.

Gay, T. Effect of speaking rate on vowel formant transitions. *Journal of the Acoustical Society of America*, 1978, *63*, 223-230.

Gay, T., & Hirose, H. Effect of speaking rate on labial consonant production. *Phonetica*, 1973, *27*, 44-56.

Gay, T., Ushijima, T., Hirose, H., & Cooper, F. S. Effect of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics*, 1974, *2*, 47-63.

Halle, M., Hughes, G. W., & Radley, J. P. A. Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, 1957, *29*, 107-116.

Klatt, D. H. Linguistic uses of segmental duration in English. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Kunisaki, O., & Fujisaki, H. On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin* (Research Institute of Logopedics and Phoniatrics, University of Tokyo), 1977, *11*, 85-91.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, 1956, *52*, 127-137.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Peterson, G. E., & Lehiste, I. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 1960, *32*, 693-703.

Pickett, J. M., & Decker, L. R. Time factors in perception of a double consonant. *Language and Speech*, 1960, *3*, 11-17.

Raphael, L. J. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America*, 1972, *51*, 1296-1303.

Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception & Psychophysics*, 1978, *24*, 471-485.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 621-637.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, 1976, *60*, 198-212.

## FOOTNOTES

[1] Since the starting and terminating formant-frequency values of the transition segment were kept constant as its duration was changed, its rate was necessarily changed as well. That is, as transition duration varied from short ([ba]) to long ([wa]), transition rate varied from fast to slow. Given that Liberman et al. (1956) have shown that transition duration, and not rate, appears to be the effective cue for the [b-w] contrast, we will refer to the stimulus manipulation in our experiments as one of duration.

[2] Summerfield (Note 2) has reported a similar finding for the syllable-initial voiced-voiceless boundary as cued by voice-onset-time (VOT). Specifically, he found that the boundary was shifted toward a longer VOT value as the syllable was lengthened by extending the steady-state vowel, but that it was shifted toward a shorter value when the syllable was lengthened by adding a final fricative.

[3] We should point out a difference between the experiments of Mermelstein et al. (Note 11) and Raphael et al. (Note 12) on the one hand, and those conducted by us on the other hand. In their experiments, the transitional information was added to the beginning of the syllable (thus, for example, changing [ɛd] vs. [ɛt] to [dɛd] vs. [dɛt], whereas we added the consonantal transitions to the end of the syllable (so that [ba] vs. [wa] became [bad] vs. [wad]). Although unlikely, it may be that the added transitions functioned differently in their experiments and ours because of the differing syllable locations to which the transitions were added.

# PERCEPTUAL EQUIVALENCE OF TWO ACOUSTIC CUES FOR STOP-CONSONANT MANNER*

H. L. Fitch,[+] T. Halwes, D. M. Erickson, and A. M. Liberman[++]

*Abstract*. First, we studied the effects on phonetic identification of orthogonal variation in two acoustic cues that are the common products of a single phonetically significant act. One component of that act produces a temporal cue, another a spectral cue. Within limits, contrasting phonetic identifications could be produced by varying either one. To determine if the implied perceptual equivalence was genuine, we measured the discriminability of stimuli in which a phonetic percept is achieved by pairing the cues in different ways. Such stimuli did prove hard to discriminate. We suggest that the equivalence thus demonstrated comes about because the two cues are processed by a system specialized to take account of their common origin in speech production. So interpreted, the equivalence can be viewed as an instance of distinctively phonetic perception, and might prove useful in experiments designed to permit comparisons among human adults, human infants, and nonhuman animals.

## INTRODUCTION

In speech, the many-to-one relationship between stimulus and percept has two aspects: Several phonetic contrasts can be produced by the same acoustic cue; conversely, several acoustic cues can produce the same phonetic contrast. Examining the first aspect, we find that the effect pervades all three phonetic dimensions. Thus, with all else constant, duration of (intersyllablic) silence, for example, can cue contrasts in manner (e.g., [ətʃa] vs. [əʃa], Note 1; Kuipers, Note 2), voicing (e.g., *ruby* vs. *rupee*, Lisker, 1957), and place (*rabid* vs. *ratted*, Port, 1976). As for the other aspect, the various acoustic cues for a particular contrast can be quite radically different. For example, an intervocalic voicing contrast in disyllables with trochaic stress (*rapid* vs. *rabid*) can be cued, all else constant, by the

*Throughout this paper, all occurrences of [ɪ] will appear as [i]. Therefore, e.g., read [slit] as [slɪt].

[+]Also University of Connecticut

[++]Also University of Connecticut and Yale University

*Acknowledgment*: This work was supported by NICHD Grant HD01994. Special thanks are due to A. Quentin Summerfield for extensive discussions in the early stages of this project and for his suggestion of the procedure used in Experiment II. We also thank Bruno Repp and David Isenberg for helpful comments on an earlier draft of this paper. Some of the work reported in this paper was presented at the Spring 1977 meeting of the Acoustical Society of America, Pennsylvania State University, University Park, Pennsylvania 6-10 June 1977.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

duration of the intersyllabic silence or, alternatively, by the formant transitions at the end of the first syllable and at the beginning of the next (Lisker, Note 3).

It is with the second aspect of the many-to-one relationship that we will be concerned. Specifically, we will put our attention on two of the cues for the manner contrast exemplified by the words _slit_ and _split_. One cue is temporal (the duration of silence between the noise associated with the initial 's' and the vocalic portion of the syllable); the other is spectral (the presence or absence of appropriate formant transitions at the onset of the vocalic portion of the syllable). Our aim is to see whether, and in what way, two such different acoustic cues can be equivalent in speech perception, and also to discover how they might be combined so as to cause their effects to summate or to cancel each other. Taken together, the results may reflect an instance of phonetic (as distinct from auditory) perception. If so, they will be the more interesting because, as we shall see, they might serve as a basis for tests of phonetic perception in nonhuman animals and human infants.

## EXPERIMENT I

The importance of silence as a cue for the perception of stop-consonant manner was shown in early studies by Bastian (Note 4) and Bastian, Delattre, and Liberman (1959). Starting with magnetic-tape recordings of the real speech utterance _sag_, the latter investigators inserted snippets of blank tape (hence silence) between the noise associated with the initial fricative and the vocalic portion of the syllable. Listeners perceived _sag_ or _stag_ depending on the duration of the (intra-syllabic) silent interval. Further studies carried out at about the same time on the contrast between _slit_ and _split_ found the perceived manner distinction to be quite categorical, not only when measured in the standard way by the relationship between identification and discrimination (Bastian, Eimas, & Liberman, 1961), but also as shown by discontinuities in the productions of subjects who tried to mimic the continuous variations of the experimental stimuli (Harris, Bastian, & Liberman, 1961).

More recent experiments on silence as a cue for manner have been designed to test the hypothesis that its effects in speech are instances of distinctively phonetic perception. On that hypothesis, one supposes that silence leads to the perception of a stop consonant, not only because the ear hears, but also, and crucially, because the silence specifies to an appropriately specialized perceptual system that the speaker has closed his vocal tract, as he must when he produces a stop. Among the data relevant to that hypothesis are some that imply a trading relation--hence an equivalence in perception--between silence and various aspects of sound that are also related to the closing (and opening) of the vocal tract. An especially telling set of such data is owing to Bailey and Summerfield (1978). These investigators found that the amount of silence necessary to produce the stop consonant in fricative-stop-vowel syllables varied with another acoustic correlate of closure--namely, the onset frequency of the first-formant. As the onset was lower, less silence was necessary to hear the stop.

In the case just described--and in other cases not involving the silence cue (see Liberman & Studdert-Kennedy, 1977, for a review)--diverse acoustic events appear to sound alike. One asks why. An answer is to be found, perhaps, in the fact that these acoustic cues are, in every case, the common but distributed products of the same linguistically significant act. Consider again, for example, the equivalence between silence and the starting frequency of the first formant in the experiment by Bailey and Summerfield. As those investigators point out, the silence occurs as a consequence of the vocal-tract closure necessary for the stop, and the low-frequency onset of the first formant occurs as a consequence of the subsequent opening of the tract. If we assume a perceiver sensitive to these cues as information about the source of phonetically significant acts, then silence and the starting frequency of the first formant might lead to the same percept because they specify the same phonetic act. On that interpretation, the trading relations we have described would, indeed, be consequences of phonetic perception.

The first of the two experiments to be reported in this paper is similar to that of Bailey and Summerfield, already described, in that it will examine a trading relation between silence and spectrum in the perception of stop-consonant manner in an initial fricative-stop cluster. Our aim is to provide a further study of this kind of trading relation, but also--indeed, especially--to obtain data that will enable us to develop a possibly useful test of phonetic perception (to be explored in our second experiment).

## Method

The stimuli of this experiment consisted of a synthetic s-like noise followed, after a variable amount of silence, by either of two synthetic syllables--one biased toward [lit], the other biased toward [plit]. Sample tokens are shown in Figure 1. To make the [lit] and [plit] stimuli, we copied appropriate aspects of the spectra of these syllables spoken by an American male and used the information to control the parameters of a serial resonance synthesizer (OVE III). Having produced synthetic approximations of [lit] and [plit], we adjusted the synthesizer parameter values so that the two syllables were identical except for the beginnings of the first three formants (F1, F2, F3). The frequency contours of the formant onsets were relatively flat for the pattern biased toward [lit] and rising for the pattern biased toward [plit], as shown in Figure 1. In addition to the intended difference in the frequency patterns of the formant transitions, the amplitude rise-time of the [lit] syllable was slightly faster than that for the [plit] syllable, a difference that was due to the particular characteristics of the OVE III synthesizer. (If the difference in rise-time had an effect, it would be to bias the stimuli *against* the trading relation we expected to find.) Each [lit] and [plit] syllable was 170 msec long.

To determine how successful we were in biasing those syllables towards [lit] and [plit], we tested the syllables, *without* any preceding s-like noise, in forced-choice identification trials. Each of five subjects heard a randomized series containing thirty instances of each of the two syllables, labeling each token as *lit* or *plit*. On the average, the stimulus biased toward [lit] was heard as [lit] 43% of the time and the stimulus biased toward [plit] was heard as [plit] 97% of the time.

To produce the full stimulus patterns, we generated a 96-msec patch of band-limited noise appropriate for [s] (referred to hereafter as 's'), placed

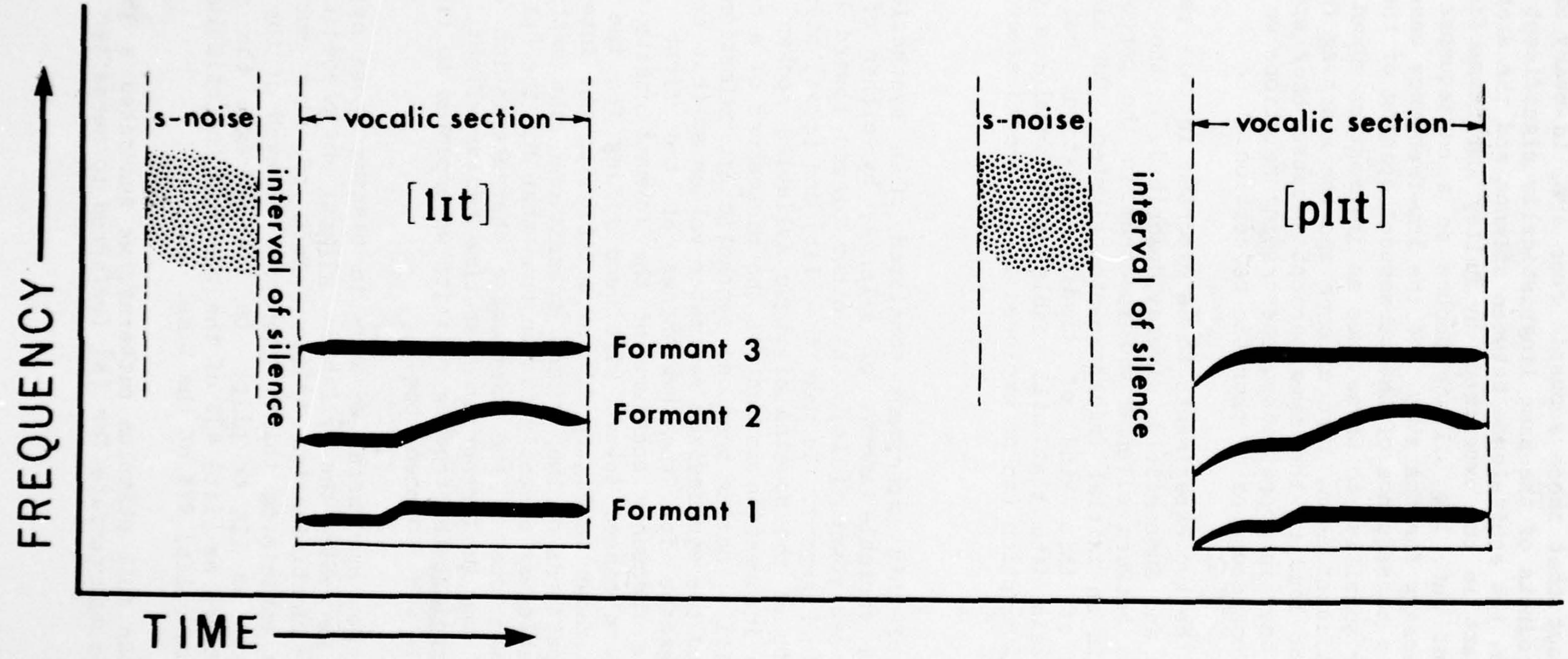


Figure 1. Schematic spectrograms of the stimulus patterns, showing two of the settings of the silent interval and both settings of the formant transitions at the onset of the vocalic portion.

it in front of the [lit] and [plit] syllables (hereafter, the vocalic portion), and varied the interval of silence between the 's' and the vocalic portion from 8 to 160 msec in steps of 8 msec, making a total of 20 stimuli in each series.

(We must emphasize that, for our purposes, not just any tokens of [lit] and [plit] will serve. Pilot work showed, for example, that some tokens of [lit]--especially those with prolonged initial transitions or very gradual amplitude rise times--would not produce [split], no matter how long the silent interval. In making our stimuli, we therefore undertook to neutralize all the acoustic parameters relevant to the [slit]-[split] contrast, save those differences of interest to us: to wit, the duration of the silent interval and the formant transitions in the initial portions of the [lit] or [plit] syllable.)

To determine the location of the phonetic boundary in each series--and thus to see the trading relation, if any, between the two cues (silence and formant transitions)--we recorded the stimuli onto a tape appropriate for presentation to listeners. That tape contained 6 randomizations of the full set of 40 stimuli, 20 from the 's'+ [lit] series and 20 from the 's'+ [plit] series. There was a 3-second pause between items and a 10-second pause between randomizations.

The experimental tape was presented once to each subject, with instructions to identify each stimulus as *slit* or *split*, and to guess if necessary. It was played over a loudspeaker at a comfortable listening level. The subjects were twelve college students. All were native American English speakers and claimed to have good hearing in both ears.

## Results

The variation in silence duration was effective in producing a perceived contrast between [slit] and [split] as is apparent in Figure 2. There we see that for both series, ('s'+ [lit] and 's'+ [plit]), all judgments shift from [slit] to [split] as the silent interval increases. More interesting is the displacement of the perceptual boundary between [slit] and [split] in the two series, for that reflects the trading relation between the silence and the spectral cue--the subject of this investigation. In that connection, we see in Figure 2 that, for the series 's'+ [lit], the phonetic boundary (here defined as the point where the interpolated function crosses the 50% level) is at about 80 msec of silence, while for the series 's'+ [plit] the boundary is at about 55 msec. Thus, it appears that about 25 msec less silence was required, on the average, to hear [split] when the formant transitions appropriate for [p] were present ('s'+ [plit] series) than when they were absent ('s'+ [lit] series). That finding defines a trading relation between the temporal cue and the spectral cue. Within the limits of that relation, these two very different acoustic cues have equivalent effects in perception.

The results we have discussed were averaged across subjects. It is appropriate, therefore, to note that, although individual listeners differed in the absolute position of the phonetic boundary, every one of the twelve manifested a boundary shift in the same direction (though not necessarily by the same amount) as for the group as a whole. The smallest shift shown by any

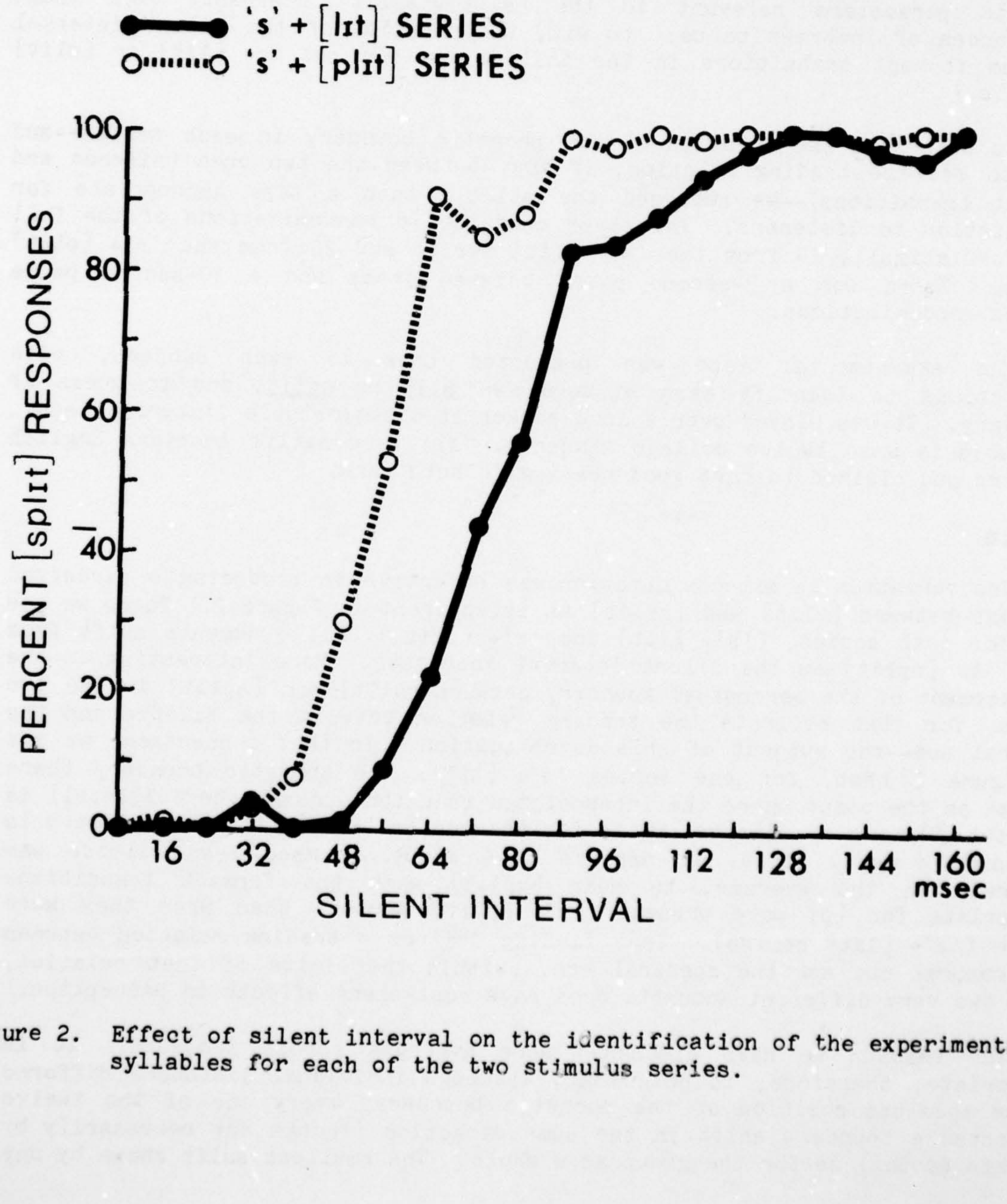
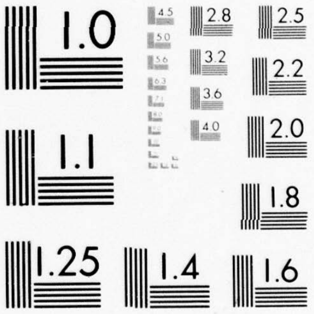


Figure 2. Effect of silent interval on the identification of the experimental syllables for each of the two stimulus series.

AD-A083 179 HASKINS LABS INC NEW HAVEN CONN F/G 5/10
SPEECH RESEARCH, A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
MAR 79 A S ABRAMSON, T BAER, C BEST, G CARDEN PHS-HD-01994
UNCLASSIFIED SR-57(1979) NL

3 OF 4
AD
A083179



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

subject was 8 msec; the largest was 40 msec.

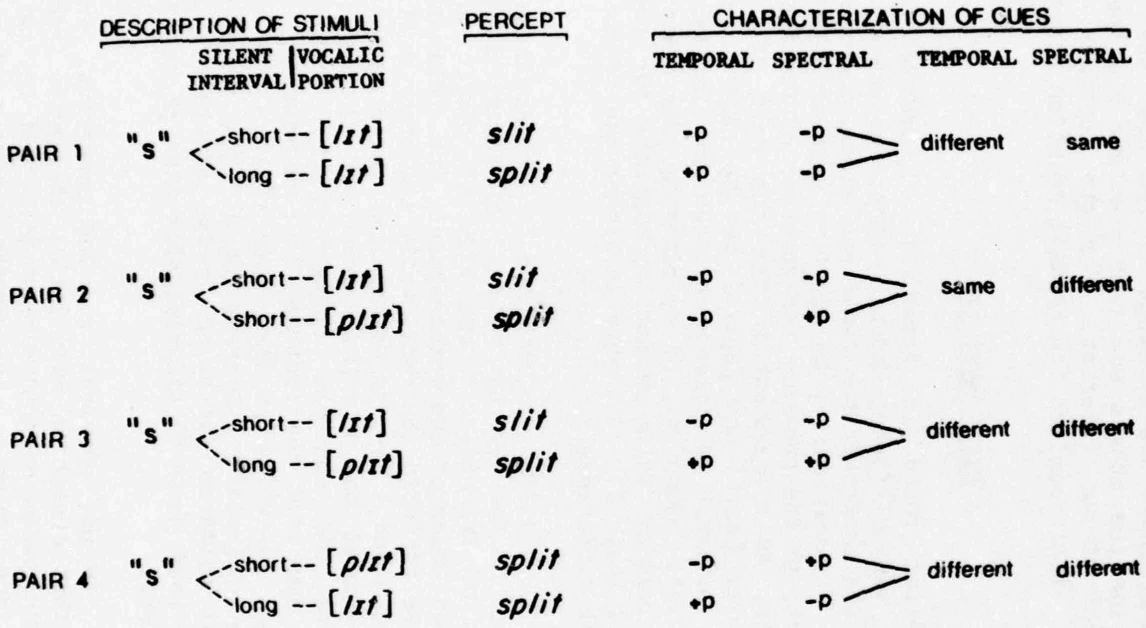
We should note further that the magnitude of the phonetic boundary shift, whether for individuals or for the group, is presumably not fixed; settings of the formant transitions other than the two we selected would likely produce a smaller or larger shift, as would different settings of other relevant cues. These latter include, for example, the offset characteristics of the 's' (Bailey, Summerfield & Dorman, personal communication). Moreover, as we observed under Procedure, there are settings of the cues that cause one or more of them to override the others and thus preclude the trading relation that is here the object of our interest. For our purposes, however, the important fact is that, within limits, the trading relation reported here does exist.

## EXPERIMENT II

The trading relation found in Experiment I implied a perceptual equivalence between two very different acoustic cues; one was silence, and its dimension of variation was temporal; the other was sound, and its dimension of variation was spectral. Specifically, twenty-five msec of silence was equal, in its perceptual effect, to the presence (or absence) of formant transitions. It follows, then, that, as diagrammed in the top half of Figure 3, the perceived contrast between [slit] and [split] can be made in either of two ways: by varying the interval of silence between the initial 's' and the vocalic part of the syllable (Pair 1), or by altering the transitions at the onset of the vocalic portion in a manner appropriate for the distinction between [lit] and [plit] (Pair 2). A change in either one of the two cues can determine the perceptual difference.

But what of pairs that differ in both cues? For an answer, we should examine the pairs shown in the the bottom half of Figure 3, where we have inferred from the data of Experiment I the kind of result that is to be expected for each of two cases of cue combination: One combination supports, or possibly enhances, the phonetic contrast, while the other nullifies it. As shown, the pairs labeled 3 and 4 have in common that their members differ from each other by both temporal and spectral cues, yet in one (Pair 3) the listener will perceive the contrast between [slit] and [split]--more compellingly, perhaps, than when the contrast is supported by either of the cues alone (Pairs 1 and 2)--while in the other (Pair 4) he will hear two noncontrasting versions of [split]. Considering all the pairs shown in the figure, we observe that patterns differing by two equivalent cues could be either more (Pair 3) or less (Pair 4) discriminable than pairs (1 and 2) that differ by either cue alone.

These effects are, to us, provocative. The perceptual equivalence is between acoustic cues that appear to have little in common from an auditory point of view. As in the cases of equivalence described in the Introduction to Experiment I, they are, however, the common products of a phonetically significant act. Perhaps, then, the equivalence is a reflection of processes specialized to perceive that common basis; if so, the equivalence might stand as an instance of phonetic perception.



| DESCRIPTION OF STIMULI | SILENT INTERVAL | VOCALIC PORTION | PERCEPT | CHARACTERIZATION OF CUES TEMPORAL | SPECTRAL | TEMPORAL | SPECTRAL |
|---|---|---|---|---|---|---|---|
| PAIR 1 "s" | short | [ɪt] | *slit* | -p | -p | different | same |
| | long | [ɪt] | *split* | +p | -p | | |
| PAIR 2 "s" | short | [ɪt] | *slit* | -p | -p | same | different |
| | short | [plɪt] | *split* | -p | +p | | |
| PAIR 3 "s" | short | [ɪt] | *slit* | -p | -p | different | different |
| | long | [plɪt] | *split* | +p | +p | | |
| PAIR 4 "s" | short | [plɪt] | *split* | -p | +p | different | different |
| | long | [ɪt] | *split* | +p | -p | | |

Figure 3. Diagrams illustrating the phonetically equivalent effects of spectral and temporal cues and the phonetically different effects of combining these cues in two ways.

The perceptual equivalence of the cues, together with its implications, is brought more compellingly to our attention by the opposite consequences of the two ways of combining them--to augment the perceived contrast or to nullify it. These consequences imply that the two kinds of cues have equivalent effects in determining the presence or absence of a single percept, [p], an implication that is most simply comprehended in phonetic terms. This is especially intriguing because, as we shall see, it may provide tests of phonetic perception simple enough (at least in principle) for use with nonhuman animals, and with prelinguistic human infants.

Having thus found these effects interesting, we should inquire into them more deeply. Specifically, we should take into account that they rest on data obtained from an experiment in which the listeners were forced to choose between [slit] and [split]. It was from those data alone that we proceeded to the conclusions captured in Figure 3. In that light, consider the conclusion that the spectral and temporal cues can work in either of two ways: cooperatively to produce (and presumably enhance) the distinction between [slit] and [split], as in Pair 3, or at cross purposes, as in Pair 4, to yield two tokens of [split]. Note, however, that the forced-choice identifications do not rule out the possibility that the listeners might have perceived a considerable difference between the two tokens of [split] as represented in Pair 4, but, lacking alternative responses, could only identify both as representatives of the same phonetic type. Hence, we cannot be sure that the members of that pair are, in fact, less discriminable than the members of the pair (3) in which the cues 'cooperate'; nor can we be sure, indeed, that the members of those two pairs are less and more discriminable, respectively, than the pairs (1 and 2) that differ by either cue alone.

Nevertheless, we have two bases for supposing that the relative discriminability of the pairs is just what the forced-choice identifications have led us to suppose it ought to be. The first is in the impressions that we, the experimenters, had in listening to the stimuli. Our impressions were that the forced-choice identification data obtained from our subjects have not obscured differences in the order of discriminability that direct tests of discrimination would have revealed. The second basis is in the data of a pilot experiment in which we presented some of the critical pairs for direct comparison, asking the (four) subjects to tell us if they could discriminate the members of the pairs. In the event, the conclusions about relative discriminability that we drew from the identification data held up well. It is fitting, however, to test those conclusions further by obtaining direct measures of discriminability in a thorough and systematic way.[1] To that end, we have carried out the experiment to be reported now.

## Method

Using the stimuli of Experiment I, we undertook in Experiment II to test the discriminability of the stimuli in three conditions. In the first, there was a one-cue (spectral) difference between the stimuli to be discriminated;

[1] Procedures appropriate for that purpose were suggested to us by A. Quentin Summerfield.

in the second and third, there were both spectral- and temporal-cue differences. The second and third conditions differed from each other in that in the second the two cues worked cooperatively, while in the third they conflicted.

In all three conditions, then, the stimuli to be discriminated differed by the spectral cue--that is, one member of each pair had a vocalic section appropriate for [lit], and the other had a vocalic section appropriate for [plit]. In the first condition, to be called the "one-cue" condition, this was the only difference; the duration of the temporal cue (interval of silence between end of 's' and beginning of vocalic section) was the same for both members of each pair. Discrimination of the spectral cue was then measured at each value of the temporal cue. Thus, at one end of the range there was a pair comprising the stimuli ('s'-8 msec silence-[lit]) vs. ('s'-8 msec silence-[plit]); at the other end, there was the pair ('s'-144 msec silence-[lit]) vs. ('s'-144 msec silence-[plit]). Between these extremes were similar pairs for all intermediate values of the temporal cue. (Since the stimuli with 144 msec silence were well past the point at which subjects heard all the tokens as [split], the stimuli with 152 msec and 160 msec of silence were omitted from the discrimination experiment in order to reduce the burden on the listeners.)

In the second condition there was not only a spectral-cue difference between the members of each pair, but also a temporal-cue difference arranged to be in harmony with the spectral cue, and thus, presumably, to facilitate the discrimination. This will be called the "two-cooperating-cues" condition. As in the one-cue condition, each pair had one member made from [lit] and one member made from [plit], but in the two-cooperating-cues condition the [plit] member of each pair had a silent interval 24 msec longer than the [lit] member. Thus, at the one end of the continuum of silent intervals, the stimulus ('s'-8 msec silence-[lit]) was paired with ('s'-32 msec silence-[plit]), and similar pairings were arranged through the entire range of silent intervals up to the pair ('s'-120 msec silence-[lit]) vs. ('s'-144 msec silence-[plit]).

In the third condition there were, again, temporal as well as spectral differences between the pairs to be discriminated, but here the cues were in conflict. We therefore called this the "two-conflicting-cues" condition. For each pair the [plit] member had 24 msec <u>less</u> silence than its [lit] companion. Thus, at the one end of the continuum of silent intervals, ('s'-[illegible] msec silence-[lit]) was paired with ('s'-8 msec silence-[plit]), and similar pairings were made at all increasing values of the silent interval through the pair ('s'-144 msec silence-[lit]) vs. ('s'-120 msec silence-[plit]).

For the two-cooperating-cues and the two-conflicting-cues conditions, choosing the amount of silence by which the members of each pair differ is, of course, critical, if the experiment is to reveal most sensitively such differences in discriminability as there may be between the two conditions, and, indeed, between either of them and the condition in which the members of each pair differed only by the spectral cue. That amount of silence would be equal to the amount by which the two perceptual identification functions-- the one for stimuli made of [lit], the other for stimuli made of [plit]-- are displaced (as in Figure 2 of Experiment I), since that is the amount of silence that, according to identification judgments, just compensates for bilabial transitions. Ideally, the amount of silence would be adjusted appropriately for each subject. For experimental convenience, we did not make the adjustment for each subject, but rather used for all a single value, 24 msec, which is close to the average displacement obtained in Experiment I.

To measure discriminability of the members of these pairs, we used an oddity test. On each trial, one member of a pair was presented twice and the other member once. The listener was to determine which of the three stimuli was the odd one. For each pair of stimuli to be tested, six different oddity triplets--that is, all possible permutations--were generated. Each of these triplets occurred three times, yielding 18 presentations per pair for the subject to judge. The full test was a random ordering of all the triplets in the experiment. The discrimination test was administered in four one-hour sessions.

At each of the four discrimination-testing sessions described above, we also obtained identification functions just as we had in Experiment I, presenting the stimuli one at a time and in random order for judgment as slit or split. This was done at the beginning and end of each session, each time obtaining 3 judgments per stimulus for each subject (altogether, 24 judgments per stimulus for each subject). The purpose of this part of the procedure was two-fold: to see if the subjects in this experiment showed the same trading relation we had found in Experiment I, and to see if the trading relation was stable across the experimental sessions. The latter purpose was especially important, since we should hesitate to combine the discrimination data across the four experimental sessions if, for any reason, the trading relation that those data are supposed to test had itself undergone some change.

Five American-English-speaking college students with no known hearing deficit served as subjects.

## Results

Part of this experiment--the identification of the stimuli as [slit] or [split]--was identical with Experiment I, though with more repetitions and with different subjects. Looking at Figure 4, we see that the same result was obtained: less silence was necessary to hear [split] when the transitions appropriate for [plit] were present. In the first experiment the average difference was 25 msec; here it was 28 msec. As in Experiment I, every subject showed a shift in the phoneme boundary and in the same direction. The smallest difference between 's'+ [lit] and 's'+ [plit] crossover points for any subject was 20 msec; the largest difference was 33 msec. No significant changes in the positions of these phonetic boundaries were noted across the four days of testing.

The results of the discrimination tests can be seen in Figure 5. Let us look first at the one-cue condition--that is, the condition in which the pairs of stimuli to be discriminated differed only in the spectral cue at the beginning of the vocalic section. Examining the appropriate data, which are shown as the solid line, we see that discrimination is relatively low at the extremes and relatively high in the region of the phonetic boundary. Given the outcome of the identification test, this is what we should have expected:

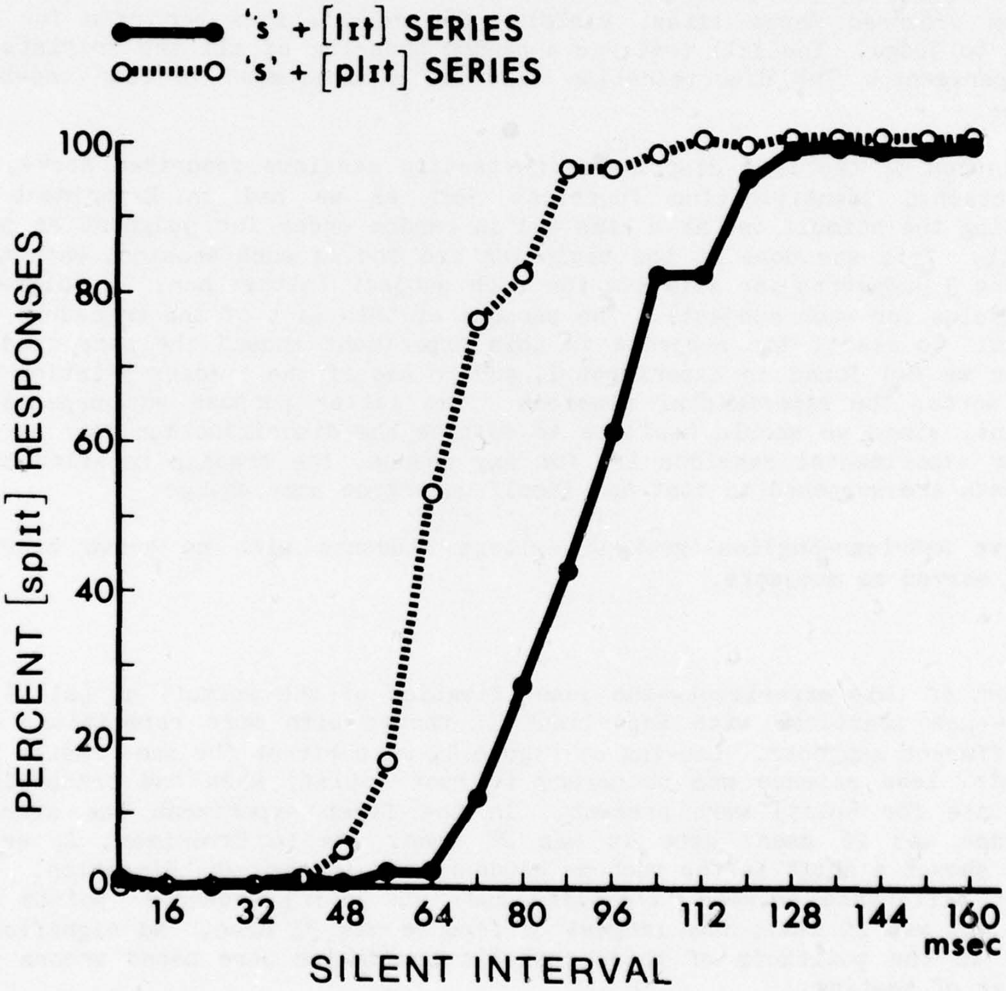


Figure 4. Effect of silent interval on the identification of the experimental syllables for each of the two stimulus series in Experiment II.

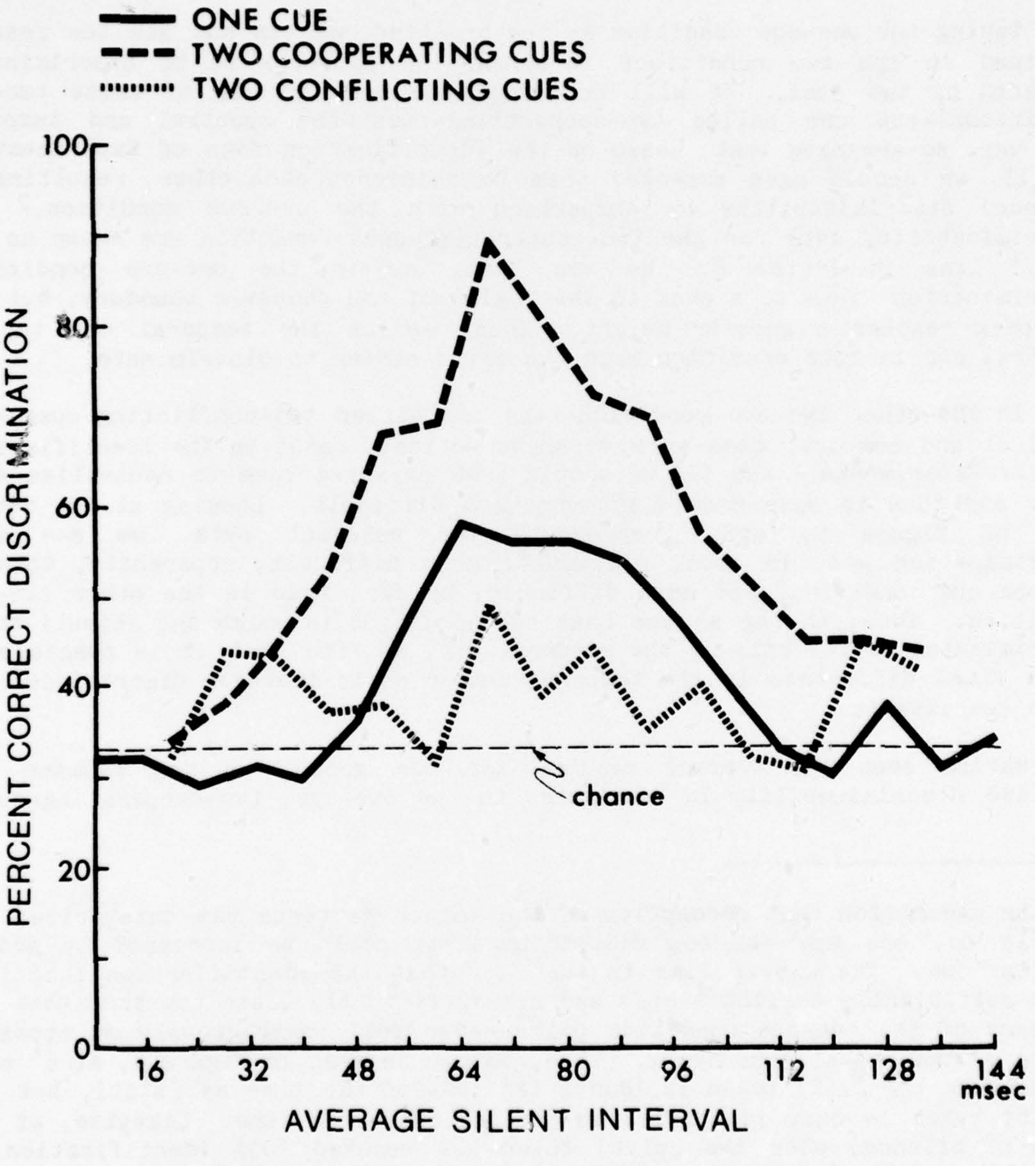


Figure 5. Percent correct discrimination for three types of stimulus pairs.

at very short intervals of silence, patterns made of 's' plus either [lit] or [plit] should have sounded like [slit]; at long intervals of silence both should have sounded like [split]; and only in the region of the phonetic boundary should there have been a relatively high level of discrimination, for there the 's' plus [plit] would have begun to sound like [split], but the 's' plus [lit] would not.

Taking the one-cue condition as the baseline, we can now see the results obtained in the two conditions in which the patterns to be discriminated differed by two cues. It will be remembered that in one of these two-cue conditions--the one called two-cooperating-cues--the spectral and temporal cues were so arranged that, based on the identification data of Experiments I and II, we should have expected them to reinforce each other, resulting in enhanced discriminability by comparison with the one-cue condition.[2] The discriminability data for the two-cooperating-cues condition are shown as the dashed line in Figure 5. We see that, as in the one-cue condition, discrimination rises to a peak in the region of the phonetic boundary, but now the peak reaches a greater height. Thus, adding the temporal cue to the spectral cue in this condition made the pairs easier to discriminate.

In the other two-cue condition--the one called two-conflicting-cues--the spectral and temporal cues were arranged so that, based on the identification data of Experiments I and II, we should have expected them to neutralize each other and thus to have made discrimination difficult. Looking at the dotted line of Figure 5, which represents the relevant data, we see that discrimination was, in fact, difficult; more difficult, apparently, than in the one-cue condition, and more difficult, by far, than in the other two-cue condition. Thus, taking as the base the condition in which the stimuli to be discriminated differ only by the spectral cue, we find that it is possible to add a fixed difference in the temporal cue so as to increase discriminability or to decrease it.

Having seen the average results for the group, we now examine the relative discriminability of the pairs in the one-cue, two-cooperating-cues,

[2] On the assumption that perception of the speech patterns was categorical, or nearly so, one may ask how discriminability could be increased by adding another cue. The answer lies in the fact that the identification functions have sufficiently shallow slopes and are sufficiently close together that the members of the one-cue condition pairs never fall unambiguously on opposite sides of the phonetic boundary. Thus, as can be seen in Figure 4, at 64 msec of silence the [lit] token is identified 100% of the time as [slit], but the [plit] token is only identified as [split] 62%of the time. Likewise, at 112 msec of silence, when the [plit] token has reached 100% identification as [split], the [lit] token is only identified as [slit] 19% of the time. Thus, even if perception of the patterns were categorical, the one-cue condition would not be expected, for any of the pairs used in the experiment, to produce discrimination at the level of 100%. In the two-cooperating-cues condition, on the other hand, some of the pairs in the middle of the series comprised stimuli that the subject had consistently put into different phonetic categories; for such pairs we should expect that discrimination would be enhanced.

and two-conflicting-cues conditions for the individual subjects. Since the differences in these three conditions occur primarily at the phoneme boundary (as we would expect), and since, although the position of the phoneme boundary varies slightly from subject to subject, it is always located near the middle of the range, we have used the data taken from the middle third of the range in determining the average level of discrimination in each condition for each subject. Those results are presented in Table 1. We see that, for every subject, the order of difficulty, from easiest discrimination to most difficult, is: (1) the condition with two cooperating cues, (2) the condition with one cue, and, (3) the condition with two conflicting cues. Thus the group results, plotted in Figure 5, accurately reflect the performance of each one of the subjects.

Table 1. Percent correct discrimination averaged over the middle third of the stimulus series for each subject. Chance is 33%.

| | SUBJECTS | | | | |
|---|---|---|---|---|---|
| CONDITIONS | 1 | 2 | 3 | 4 | 5 |
| Two Cooperating Cues | .77 | .75 | .81 | .77 | .70 |
| One Cue | .56 | .55 | .56 | .50 | .46 |
| Two Conflicting Cues | .40 | .43 | .42 | .34 | .38 |

To compute the probability that this result could have been obtained by chance, we may take advantage of the fact that we expected a particular ordering of the three conditions. Since there are six possible orderings of three conditions, and we expected one of these, then for any subject the exact likelihood of this particular ordering of results in the three conditions is one in six. The likelihood of obtaining this same ordering for each of five subjects is $(1/6)^5$, or once in 7,776 experiments.

Having seen the relative order of difficulty in discriminating the different types of syllable pairs, we should recall what this has to do with the claim that the spectral and temporal cues are equivalent when they converge on the perception of [p] in [split]. For that purpose, we put our attention first on the two-conflicting-cues condition. If the two cues are truly equivalent in phonetic perception, then it should be possible to arrange them so that they effectively neutralize each other, producing two syllables--for example, [split] and [split]--that sound alike. That is, adding duration of silence to a vocalic pattern that lacks the spectral cue for [p] should produce a syllable [split] that is hard to distinguish from one in which subtracting duration of silence has been compensated for by the presence of the spectral cue for [p]. Of course, these additions and subtractions must be of appropriate and reasonable magnitudes, which is to say that they must produce a perceived contrast between [slit] and [split] in patterns that differ by only one of the cues, or by both cues when put together in the other order (two-cooperating-cues). In fact, as we have seen, the pairs formed by the two conflicting cues were relatively indistinguishable by comparison with either of the other pairs. We conclude that the two cues did effectively neutralize each other in perception. That is to say that they did, indeed, sound alike.

## DISCUSSION

As pointed out in earlier sections, the several cues for a phonetic contrast are typically found, in perceptual studies, to engage in a trading relation: given a phonetic contrast for which each of two cues is relevant, the effects of varying one cue can, within limits, be compensated for by appropriate variations in the other. Such trading relations are of interest if only because they imply an equivalence among aspects of stimulation that are often quite different from an acoustic point of view. They are the more interesting if we are right in supposing that the equivalence reflects a sensitivity to the common origin in articulation of the different acoustic cues--that is, to processes that are distinctively phonetic.

The trading relation observed in Experiment I is novel only in that it is another token of a type. Surely, our token is a striking one, for the contrast between the cues that trade is very great indeed: one of them is silence, and its dimension of variation is temporal; the other is sound, and its dimension of variation is spectral. But even that contrast is not entirely new, as the reader can find in the cited paper by Bailey and Summerfield (1978). There he will see experiments on trading relations different from ours and also considerably more comprehensive. But we need note here only that, like us, Bailey and Summerfield studied the contrast between fricative-vowel and fricative-stop-vowel syllables; they employed temporal variations in silence and spectral variations in sound, just as we did; they found trading relations not different in principle from ours; and they offered an interpretation that is, in some important respects, similar to the one we favor.

But trading relations among acoustic cues, with the phonetic equivalence they imply, have been based on perceptual tests that only require of a subject that he attach phonetic labels. As pointed out in the Introduction to Experiment II, this leaves open the possibility that, for want of an alternative, subjects might sometimes have attached the same phonetic label to stimuli that were, in some peculiar way, as different as two stimuli to which they had found it possible to assign different labels. What is novel about our experiment is that we have subjected the claim of perceptual equivalence to a more rigorous test by having our subjects discriminate the stimuli on any basis whatever. The results of that discrimination test justified the inference about equivalence that had been made from the way the subjects assigned phonetic labels.

Now we would remark a possibly interesting by-product of Experiment II: It may provide a test of phonetic perception that can be applied to nonhuman animals. Consider, again, the result with adult humans, which was that the two very different cues (silence and sound) could be so combined as to 'neutralize' each other and thus produce pairs of syllables that are hard to discriminate, harder than pairs in which the same two cues are made to augment each other, but also harder than pairs that are distinguished by one of the two cues alone. That the cues have such effects can reasonably be taken to mean that, in the proper phonetic context, they are actually equivalent--that is, they do sound alike. But if we are right in supposing that the syllables with the different cues sound alike only to animals specialized to perceive their phonetic import, then to the nonhuman animals they should sound quite different--as different, presumably, as they would to us when heard in a nonphonetic context. In that case, we should suppose that the two cues could not be made to neutralize each other. The result for the animals would then be that both of the pairs differing by the two cues would be easier than the pairs that differ by either cue alone. In any case, the outcome of the appropriate experiments could be straightforward and telling. What we should have to look for is only a difference in the relative order of discriminability among three pairs of stimuli. Moreover, the result to be expected is that the pair that is relatively the hardest for the human beings to discriminate would be one of the two easiest for the nonhuman animals. Such a result, if obtained, could not be attributed to inattention or lack of motivation.

We realize that such an experiment might prove to be difficult in practice. The experiment requires a demonstration, before the critical test, that the animals are able to discriminate each of the two cues taken singly. But the animals might well be defeated by the particular kinds of cues and contexts used in the experiments reported here. If so, we might expect, in further research on adult human beings, to find other sets of cues that satisfy our requirements. Indeed, all cues that engage in trading relations are candidates.

If, in the event, animals do show the different order of relative discriminability that we rather expect, then it would, of course, be of interest to apply the same test to human infants. Obviously, the considerations that make it a good test for animals would apply equally to human infants, though, just as obviously, it might be difficult with human infants, just as with nonhuman animals, to find cues that are discriminable yet appropriate. But if such cues can be found, then the kind of test we have proposed might provide a useful way to reveal (and study) an important biological predisposition to language.

In summary, we have two very different acoustic cues that engage in a trading relation and are, in an important sense, perceptually equivalent. We suggest that this equivalence is due to the fact that the cues are the common products of a single phonetically significant act, and are perceived by a system specialized to take account of that fact. If that is so, then a test for perceptual equivalence of two such different cues may provide an interesting basis for comparative studies among adult humans, pre-verbal infants, and non-human animals.

## REFERENCE NOTES

1. Affricates: Duration cues for the perception of č in intervocalic position. (Quarterly Progress Report, 1954, No. 12) Haskins Laboratories, New York.
2. [Kuipers, A. ]. Affricates in intervocalic position. (Quarterly Progress Report, 1955, No. 15, Appendix 6) Haskins Laboratories, New York.

3. Lisker, L. Closure duration, first-formant transitions, and the voiced-voiceless contrast of intervocalic stops. (*Quarterly Progress Report*, 1957, *No. 23, Appendix 1*) Haskins Laboratories, New York.
4. Bastian, J. Silent intervals as closure cues in the perception of stop phonemes. (*Quarterly Progress Report*, 1959, *No. 33, Appendix 1*) Haskins Laboratories, New York.

## REFERENCES

Bailey, P. J., & Summerfield, A. Q. Some observations on the perception of [s] + stop clusters. *Haskins Laboratories Status Report on Speech Research*, 1978, *SR-53 vol. 2*, 25-60.

Bastian, J., Delattre, P., & Liberman, A. M. Silent interval as a cue for the distinction between stops and semi-vowels in medial position. *Journal of the Acoustical Society of America*, 1959, *31*, 1568 (A).

Bastian, J., Eimas, P. D., & Liberman, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America*, 1961, *33*, 842 (A).

Harris, K. S., Bastian, J., & Liberman, A. M. Mimicry and the perception of phonemic contrast induced by silent interval: Electromyographic and acoustic measures. *Journal of the Acoustical Society of America*, 1961, *33*, 842 (A).

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz & H.-L. Teuber (Eds.), *Handbook of sensory physiology*, vol. VIII. Heidelberg: Springer-Verlag, 1977.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 1957, *33*, 42-49.

Port, R. Influence of tempo on the closure interval cue to the voicing and place of intervocalic stops. *Journal of the Acoustical Society of America*, 1976, *59*, S41-42 (A).

# SYLLABIC CODING AND READING ABILITY IN WORD RECOGNITION

Leonard Katz[+]

Abstract. Two experiments were run in order to determine if fifth grade children use syllabic coding as a component of the word recognition process in reading. Contrary to results from previous studies, evidence for syllabic coding was found in a lexical decision task for high frequency words and, less strongly, for low frequency words. In a naming task, an effect of syllabic coding was found for pseudoword naming latency but not for high frequency words. However, syllabic coding effects were found on error measures for both pseudowords and high frequency words. Throughout both experiments, skilled readers performed better than less skilled readers but the syllabic coding effects were similar for both reading ability levels.

## INTRODUCTION

The experiments presented here looked for evidence of effects on reading due to a major phonological variable, the syllable. Two questions were asked: (1) is the syllable a component of word recognition in reading and, if it is, (2) are skilled readers better than the others at utilizing syllabic information to recognize words?

Liberman and Shankweiler (1978) demonstrated that poor readers often fail to achieve awareness of the syllabic and phonetic structure of spoken language, an awareness that is necessary to use an alphabetic writing system effectively. How lack of awareness of the underlying structure of speech can lead to difficulties with reading is described systematically in the reading model of Laberge and Samuels (1974). (See also Estes, 1977, for a similar treatment.) Reading is seen as a process by which visual information is transformed in a series of processing stages involving visual, phonological, episodic memory and semantic memory systems. Some of the visual codes derived from print have their counterparts in phonological codes; letters, spelling-patterns and whole words in the visual system are in complex association with phonemes, syllables and words in a phonological system whose units are closely related to acoustic and articulatory inputs. The information flow is from subunits (for example, letter features) into large units (for example, letters) arranged in a hierarchical structure. For the skilled reader in the Laberge-Samuels model, reading becomes more efficient as the various sub-

[+] Also University of Connecticut.

Acknowledgment: Carol Fowler collaborated on Experiment 1 and Marsha Goldberg assisted with data analysis. I am grateful to Carol Fowler, Virginia Mann, and Rebecca Treiman for valuable criticism of an earlier draft. This research was supported by NICHD Grant HD-01994 to Haskins Laboratories.

processes require less attention, i.e., become more automatic. Efficiency requires that the appropriate information processing structures evolve, and the beginning reader who attempts to map entire printed words directly into a phonological equivalent without first establishing a substructure based on the components of the spoken word (e.g., phonemes and syllables) will not become a skilled reader. It might be the case, then, that readers who have evolved processing structures without having been aware of the existence of syllables in speech when they began to learn to read might be destined to become less efficient readers. On the other hand, the less skilled readers might eventually restructure the process of mapping print into syllable units as their awareness of the spoken language matured, in which case, no trace of the earlier structure might be found.

In the experiments reported here, the major experimental manipulation is the degree of integrity of the syllable units in each stimulus. Two aspects of word recognition are studied: pronunciation and meaning. If syllabic information is used in either aspect, then disruption of the syllable unit should prove deleterious. The first experiment uses a naming task in an attempt to bias the subject toward the use of a phonological strategy for word identification. Although, to my knowledge, syllable unity has not been manipulated before, another syllable variable--the number of syllables--has sometimes been found to affect the latency of naming a printed word or pseudoword (Ericksen, Pollack, & Montague, 1970); but negative evidence has also been found (cf. Fredriksen & Kroll, 1976). The second experiment attempts to bias the subject toward a search for the meaning of a stimulus; a lexical decision task is used. Although evidence for some kinds of phonological encoding of the stimulus has been found in the lexical decision task under certain conditions (e.g., Davelaar, Coltheart, Besner, & Jonasson, 1978), there appears to be no evidence of specific syllable effects (Fredriksen & Kroll, 1976).

In both experiments children of skilled and less skilled reading abilities are studied with regard to their use of syllabic information. Skilled readers were expected to perform better than less skilled readers. Of specific interest are the possible interactions of reader ability with syllable unity. As suggested above, it may be the case that less skilled readers do not use syllabic information in the same way as do the skilled readers.

Recently, there has been support for the notion that both phonological and nonphonological (direct visual) codes of print can access the reader's internal lexicon, i.e., his memory for words. Coltheart, Davelaar, Jonasson, and Besner (1977) and Meyer, Schvaneveldt, and Ruddy (1974) agree that both phonological and nonphonological routes to the lexicon are possible; they differ in that Coltheart et al. suggest that mutual facilitation between routes is possible, while Meyer et al. do not allow for that possibility. Marcel and Patterson (1978) present data from studies of aphasics and normals that also support the dual channel notion. Levy (1977) found severe and stable deficits in reading comprehension when vocalization was required; no such deficits occurred in analogous listening comprehension tasks. However, Levy also found comprehension deficits unrelated to vocalization (thematicity effects) and concluded that both speech processing and visual processing are used in accessing the meanings of both individual words and sentences.

Experimenters who probe for evidence of phonological encoding in reading will miss finding certain visual codes used by a reader that are related to phonological codes; these visual codes may still preserve some characteristics of the original phonological encoding. Specifically, it may be the case that visual codes of spelling patterns exist that reflect a past association in the reader's experience of these spelling patterns with historically older phonological syllable codes. It might be expected that the information in these visual codes preserves the syllabic divisions present in the phonological codes that preceded them. If so, we would find a unitization of letter groups into syllable-like visual units. This visual, but syllable-like, encoding might or might not be merely a temporary stage in the evolution of visual reading codes; they might or might not be found in mature reading behavior as well as in children's reading. Either way, it may be the case that the reader (at least at some stage of development) visually parses words into syllables (among other parsings) because his overt and covert speech systems, which were operational in earlier reading acquisition, did so previously.

The present experiments were not designed to pinpoint the locus of syllable effects in either visual or speech codes. Rather, these studies look for any evidence of syllable effects on word recognition while using a technique that presumably can probe for a reader's sensitivity to the syllables in the words he or she is reading, whether that sensitivity depends on information in the visual mode or in the speech mode. The technique requires dividing a printed multisyllabic word with a nonalphabetic marker (a slash). Two types of division are used: division of the word that corresponds to correct syllabification, and division of the word that is incorrect with regard to syllabification. The words that are correctly syllabified should "look right" to any visual information encoder presented with the printed stimulus. If a visual syllabic code is used in accessing the name or meaning of a word, then such a code will be facilitated by correct syllabification compared to an incorrect division of the word.

If instead of (or in addition to) visual processing, speech modality codes are used in word recognition, then the stimuli that are correctly syllabified should again be easier to transform from print to speech. Finally, if syllable processing is not important to word identification, it will suggest that one kind of intraword phonological processing does not affect word recognition, even through the use of a related visual code. Note that the statement that syllabic coding exists does not imply that the structures of the word naming process or the lexical memory, etc., are to be viewed as syllabaries; the only implication to be made is the weaker one: that syllabic coding is utilized somewhere in the process of recognizing words.

## EXPERIMENT 1

In order to encourage phonological processing (and therefore heighten potential syllable effects) subjects were required to respond to a stimulus item by pronouncing it (naming). Stimuli were divided (using a slash) either appropriately (Regular) or inappropriately (Irregular) and both skilled and less skilled readers were studied. In addition, word familiarity was varied; high frequency words were contrasted with orthographically and phonologically regular pseudowords.

The word-pseudoword variable was introduced to examine the possibility that high frequency words are not processed phonologically, but, rather, that their pronunciations are accessed lexically. Lexical access to pronunciation may occur in a manner similar to (and perhaps along with) access to semantics (cf. Forster & Chambers, 1973; Marcel & Patterson, 1978). However, there is no such memory of a pseudoword's pronunciation; subjects are likely to pronounce them by applying phonological rules. Therefore, if one expects to find a syllable effect anywhere, it should occur at least with pseudowords. Nevertheless, it should be noted that subjects are not limited entirely to phonological processing in order to pronounce a pseudoword, because pronunciation can still take place, in part, by finding real word analogs to the printed pseudowords (or analogs to parts of the pseudoword) and lexically accessing the pronunciations of those analogs.

## Method

Subjects. From a group of 51 fifth grade children, 25 skilled readers and 26 less skilled readers were chosen. The mean reading class levels on a scale of 0 - 9 and mean Comprehensive Test of Basic Skills reading grade equivalence were 6.6 and 8.7 for the skilled readers, and 3.2 and 5.1 for the less skilled. The children were tested in October and November of the school year and had been given the CTBS the previous June.

Stimuli. Forty-five high frequency two-syllable words were chosen from the 1,000 most frequent words in the Carroll, Davies and Richman (1971) norms. Each of forty-five pseudowords was constructed by first permuting the vowels of a high frequency real word. Often, this produced a pseudoword that was phonologically irregular or very unusual, and a new vowel was used instead. Examples of real words and their pseudoword counterparts are: letter -lutter, region - rogion, among - omang, coming - cimong, perhaps - parheps.

Each word was syllabified to produce a Regular and an Irregular version. Each word was syllabified appropriately using either a standard dictionary or the rules suggested by Spoehr and Smith (1973). Each word was also divided inappropriately at a position that conformed to no standard rules of correct syllabification. The position of each appropriate division was matched by an inappropriate division of some word; i.e., the position of the division was balanced. A division was marked with an oblique slash that occupied one character space. The stimuli were placed on cards 10.16 x 15.24 cm using Chartpak Alternate Gothic 2 lowercase type. A quasi-random ordering of 90 Regular and Irregular stimuli was produced for words and an homologous one for pseudowords. If a Regular stimulus appeared in the first 45 trials, its Irregular counterpart appeared in the second 45 trials, and vice versa.

Design. A between-subjects design was used, with 11 skilled readers and 13 less skilled readers receiving real words, and 14 skilled and 13 less skilled readers receiving the nonword condition. Because of machine failure, one additional skilled reader was discarded.

Procedure. Subjects were run on a Gebrands three-channel tachistoscope. A fixation frame came on for 800 msec, followed by the stimulus for 1.5 sec, and a blank frame for 1 sec. Stimuli subtended approximately 2 degrees of visual angle. Subjects were told to say each word quickly when it appeared.

A voice operated relay was adjusted during five practice trials and then used to clock the interval between stimulus onset and the onset of vocalization. After practice, two dummy trials preceded, without a break, a run of 45 trials. There was a brief rest after trial 45 while the cards were changed for the second run of 45 trials. The interstimulus interval was controlled by the experimenter and was approximately 12 seconds.

## Results and Discussion

Errors were classified into one of two types: incomplete utterances (e.g., partial vocalizations) and mispronunciations. Incomplete utterances were scored as such only when followed immediately by a correct pronunciation. Long hesitations between the two syllables were also scored as incomplete. With regard to scoring mispronunciations, the experimenter's criterion for a correct pronunciation was generous for both real words and pseudowords.

An analysis of variance was performed on the number of incomplete utterances, for each subject. The effect of reading ability was significant, with skilled readers making only 1.04 errors while the less skilled readers averaged 3.21 errors, $F(1,47) = 6.06$, $MS_e = 6.67$, $p = .018$. No other effects involving reading ability were significant. Fewer errors were made on real words than on pseudowords, $F(1,47) = 13.17$, $MS_e = 6.67$, $p < .001$, and fewer errors were made on regularly syllabified stimuli than on on irregular stimuli, $F(1,47) = 6.05$, $MS_e = 4.29$, $p < .017$. The interaction of these two variables was also significant, $F(1,47) = 4.05$, $MS_e = 4.29$, $p < .05$, and is given in Table 1. Clearly, the disruptive effect of irregular syllabification was stronger for pseudowords than for high frequency real words, although a syllable effect exists for both.

TABLE 1: Errors (mean number of incomplete utterances and mispronunciations) and naming latencies (in msec) for correct pronunciations (Experiment 1).

| | Real Words | | Pseudowords | |
|---|---|---|---|---|
| Stimulus Division | Regular | Irregular | Regular | Irregular |
| Incomplete Utterances | .65 | .87 | 1.69 | 3.53 |
| Mispronunciations | .61 | 1.07 | 1.90 | 3.22 |
| Latencies | 655 | 656 | 1007 | 1055 |

The results were slightly different for mispronunciations. There was no significant difference between skilled and less skilled readers on these errors, skilled readers making an average of 1.43 errors while less skilled readers made 1.96 errors. As above (see Table 1), there were significant effects for the word-pseudoword comparison $F(1,47) = 14.47$, $MS_e = 75.03$, $p < .001$ and the regular-irregular syllabification comparison, $F(1,47) = 8.65$, $MS_e = 20.17$, $p < .006$, but their interaction was not significant, unlike the result for incomplete utterances. The significant interaction for incomplete utterances suggests that syllabification is more important for naming novel stimuli than for naming frequent words. Although the interaction leading to this suggestion is only marginally significant, the same interpretation is supported more strongly by the response latency data.

Mean latencies were computed for correct responses on regular stimulus trials and irregular stimulus trials for each subject. An analysis of variance performed on mean latency produced a pattern of results similar to the analysis performed on incomplete utterances, but more highly significant. The mean latencies, in msec, were 771 for skilled readers and 915 for less skilled readers, $F(1,47) = 13.23$, $MS_e = 39784$, $p < .001$. No other effects involving reading ability were significant. Table 1 presents the mean latencies for regularly and irregularly syllabified words and pseudowords, averaged over both reading ability groups. Not surprisingly, real words were pronounced faster than pseudowords, $F(1,47) = 89.53$, $MS_e = 39784$, $p < .001$, and the effect of regularity is also significant, $F(1,47) = 13.28$, $MS_e = 1148$, $p < .001$. However, the interaction of the two variables clearly accounts for the main effect of regularity; irregular syllabification is detrimental only when pronouncing pseudowords, $F(1,47) = 12.32$, $MS_e = 1148$, $p < .001$. The finding of a syllable effect on the naming of pseudowords is striking in the light of negative evidence for syllable involvement in naming either words or pseudowords from Henderson, Coltheart, and Woodhouse (1973), Forster and Chambers (1973) and Fredriksen and Kroll (1976). However, these investigations varied the number of syllables in the stimulus rather than the integrity of the syllable unit and that difference with the present experiment may be the critical one. One possibility is that phonological processing of pseudoword (or novel) stimuli occurs, but such processing is unlike inner speech in that it is not necessarily distributed in time; the various syllables are processed in parallel. The technique of syllable disruption used in the present experiment, however, may delay the onset of that phonological processing that is dependent on a syllabic partitioning of the stimulus. The data are clear with respect to the effect of syllable regularity on pseudoword naming; both errors and latencies indicate that regularly syllabified stimuli are easier to process than irregular stimuli. The results for real words are contradictory; syllable effects are indicated by both types of error responses but not by latencies. While it is possible that subjects were attempting to maintain a constant response latency and were trading-off speed against errors, this hypothesis does not gain plausibility from the low error rates.

With respect to reader ability differences in the processing of syllable information, the experiment found no interactions between reader ability and syllable regularity. Therefore, the substantial reader ability differences that were found appear to be unrelated to the processing of syllable information.

The appropriateness of articulation as a measure of phonological processing may be questioned. Fredriksen and Kroll (1976) accept such an equation between the phonological representation of a letter string and the articulatory representation of the string, but Davelaar et al. (1978) do not. Fredriksen and Kroll assume that, in word naming, phonological encoding is completed first and that this encoding then determines the articulatory coding for the response. Davelaar et al. point out that the articulation of a letter string may, in fact, begin before the entire string is phonologically encoded, and, therefore, a distinction must be made between the final phonological and articulatory codes. Forster and Chambers (1973) go somewhat further and suggest that, if a phonological code is used, the code used for naming is not necessarily the same as the phonological code used for lexical access. In the present experiment, there is some evidence to support Davelaar et al.; the incomplete utterances appear to be overt examples of articulation that began before phonological coding was completed. The likelihood of a difference between the code used for articulation and the code (possibly phonological) normally used to access the meaning of a word in natural reading is, perhaps, enhanced by the present paradigm in which rapid articulation is required, but access to meaning is not.

## EXPERIMENT 2

The second experiment was designed to require processing for meaning, thereby including an important component of natural reading. In addition, Experiment 2 extended the study of real words to include low frequency words as well as high frequency words. The finding of syllable effects on latencies for pseudowords but not for high frequency words in Experiment 1 suggested that, if a similar result were obtained in Experiment 2, low frequency words would be found to behave more like pseudowords, i.e., low frequency words would show evidence of syllable processing on latencies. A lexical decision task was chosen as the experimental paradigm. In the lexical decision task, word frequency effects are typically found and those were expected here; reaction times to high frequency words were expected to be faster than low frequency words, and words were expected to be faster than pseudowords. As in the previous experiment, I studied both skilled and less skilled reader groups. Of primary interest were the questions of (1) whether the regularity of syllabification would affect reaction times to high and low frequency words as well as to pseudowords and (2) whether these syllabification effects would interact with reader ability.

### Method

Subjects. From a group of fifth grade children, 51 skilled readers and less skilled readers were selected. From this set, 18 skilled and 18 less skilled readers were chosen for the major analyses on the basis of response error criteria discussed below. Unless noted otherwise, all remarks apply to the final group of 36 children. The children had been classified previously into one of 10 reading classes using the Comprehensive Test of Basic Skills together with teachers' recommendations. Skilled readers were selected from reading classes 7-9 and less skilled readers from the range 0-4. The mean reading class levels and CTBS reading grade equivalence were 8.56 and 10.4 for the skilled readers and 2.67 and 6.3 for the less skilled readers. The children were studied in the experiment in May and June and were given the CTBS in June. Therefore, as expected, their CTBS scores were somewhat higher than those of the children in the first experiment (run earlier in the school year) whose CTBS scores date from the end of the previous school year.

Stimuli. Six two-syllable high frequency and six two-syllable low frequency words were selected from the Carroll et al. (1971) fifth grade norms. The high frequency words had a range of from 145 to 959 occurrences in a sample of 634,283 tokens with a mean frequency of 411. The low frequency words ranged from a frequency of 1 to 10 occurrences, with a mean of 3.5. Each low frequency word was required to be similar in length to a specific high frequency word and to preserve morphemes such as -er and -ing in the high frequency word. For each high and low frequency word, a control pseudoword was constructed by changing one or more consonants or vowels in the word. Some of the words were composed of suffixes attached to stems that were themselves words (e.g., poster, wanted). For the set of low frequency words, the word frequency of the stem was always lower than the entire word; for the set of high frequency words, the stem was always more frequent than the entire word.

Pilot testing established a pool of words and pseudowords that were reliably judged correctly. In pilot tests, stimuli were printed on a sheet of paper and fifth grade poor readers were asked to circle those stimuli that were real words. The final set of stimuli is presented in Table A of the Appendix. Thus the 24 stimuli may be viewed as divided into six sets of four conditions each, each set consisting of a high frequency word, a low frequency word, and two pseudowords. Within each set, the four stimuli are structurally similar with regard to orthography and syllabification; the similarity is greater between a pseudoword and its real word counterpart. Each stimulus was syllabified once regularly and once irregularly by means of a slash, as in the previous experiments, to produce 48 stimuli.

Two lists, each containing all 24 original stimuli, were constructed. For List A, three of the six stimuli in each of the four conditions (high and low frequency words, pseudoword controls for high and low frequency words) were syllabified regularly and three were divided irregularly. Each list contained a quasi-random sequence of conditions. List B was identical to List A, except that a stimulus that was divided regularly in List A was divided irregularly in List B, and vice versa (see Table A in the Appendix). Stimuli were placed on 10 x 15 cm tachistoscope cards as in Experiment 1. As in Experiment 1, each stimulus subtended approximately 2 degrees of visual angle.

Design. Nine skilled and nine less skilled readers received List A followed by List B, while the remaining nine children in each reader ability group received the reverse order of lists. Randomized within each list were the factorial combinations of the four conditions (high and low frequency words, two pseudoword controls) by two levels of syllable regularity.

Procedure. Subjects were told to decide whether each stimulus displayed was a real word or not, based on their knowledge of the meaning. If they did not know the meaning of a stimulus, they were told to judge it as not a word. The experimenter showed the child four cards representing real and pseudowords, two divided regularly and two divided irregularly. The slash in each stimulus was explained as the experimenter's way of making the task more challenging for the child. Then four practice trials were given. The child pressed a telegraph key with his or her dominant hand to make a "word" response, the other hand being used for a "not word" response. The first list was then presented, preceded without interruption by a dummy trial. The cards were changed after the first list; and then the second list, preceded by a dummy trial, was presented.

Each trial began with an 800 msec fixation frame, followed by the stimulus frame for 1500 msec, followed by an all-white frame for 1 sec.

## Results and Discussion

Errors were frequent. They consisted mainly of wrong key presses ("word" instead of "not word" or vice versa) even though the subject often knew the correct identity of the stimulus. Such responses were usually followed immediately by the correct key press or by a remark that the subject momentarily confused which key went with which decision. A score of five errors or less (out of 48 trials) was chosen as a criterion for including a child in the final latency data analysis and subjects were run until the quota for each cell of the design was filled with criterion subjects. A total of 51 children was run in order to obtain a final set of 36 criterion children. Of the rejected subjects, six were skilled readers and nine were less skilled readers.

Error analyses were performed both on the total set of 51 children and on the criterion set of 36. Analyses on the larger set provided less clear results (e.g., less regular data, lower F-ratios) than analyses on the criterion set. For the criterion set of subjects, an analysis of variance was performed on the mean number of errors in each combination of condition (high frequency words, pseudoword controls for high frequency words, pseudoword controls for low frequency words), syllable regularity (regular, irregular), and time (first stimulus list, second list) for each subject in each reader ability group (skilled, less skilled). There were large effects of reader ability, $F(1,34) = 15.5$, $MS_e = .179$, $p < .001$. The mean total number of errors for skilled readers was 1.8 (3.7%) and for less skilled readers was 4.0 (8.3%). The effect of time was also significant, with an average of 3.7 errors the first half of the session and 2.0 errors the second half; $F(1,34) = 11.8$, $MS_e = .150$, $p < .002$. Condition was also significant, $F(3,102) = 6.12$, $MS_e = .146$, $p < .001$. The mean errors for conditions were: high frequency words, 1.44, low frequency words, 3.89, pseudoword high frequency controls, 2.22, pseudoword low frequency controls, 4.0. The effect of syllable regularity was not significant. The interaction of time, condition and regularity was the only significant interaction, $F(3,102) = 3.21$, $MS_e = .180$, $p < .03$. The interaction data appear to concern low frequency words. For that condition, subjects were more likely to make errors on regularly syllabified words than irregularly divided words in the first half of the session and then reversed that pattern for the second half. Compared to the three strong main effects that were obtained, the interaction result appears to be minor. The analysis of variance of errors based on the larger set of 51 children, which included both criterion and noncriterion subjects, gave similar results to the analysis of variance on the criterion set of subjects, with one exception: viz., the three-way interaction of time, condition and regularity was not significant. Of most interest to the present experiment is the fact that in neither analysis were the effects of syllable regularity or its interactions significant. However, both analyses produced F-ratios for the interaction of regularity and conditions that were nearly identical and approached significance ($p < .08$). These marginally nonsignificant results are of interest only because the same interaction turned out to be highly significant in the analysis of reaction times.

An analysis of variance was performed for reaction times on correct responses. As in the error analyses, the three main effects of reader ability, time, and condition were significant, while the main effect of syllable regularity was not. Only one interaction was significant, syllable regularity by condition. No other terms approached significance. The reader ability mean latencies were 900 msec for skilled readers and 1036 msec for less skilled readers, F(1,34) = 6.21, $MS_e$ = 426,433, p = .018. For the factor of time, the mean first and second half latencies were 1011 and 925 msec, respectively. Figure 1 presents the mean latencies for the combinations of syllable regularity and condition. For the condition main effect, F(3,102) = 47.47, $MS_e$ = 23,321, $p < .001$. For the interaction, F(3,102) = 6.26, $MS_e$ = 10,394, $p < .001$.

Figure 1 shows that latencies for the four conditions are ordered from fastest to slowest as follows: high frequency words, low frequency words, pseudoword controls for high frequency words and pseudoword controls for low frequency words. An interpretation of the interaction of syllable regularity and conditions appears to be straightforward upon inspection of Figure 1. For the word conditions, latencies to regularly syllabified words were faster than to irregularly divided words while for the pseudoword conditions, the reverse was the case. Moreover, it appears that the magnitude of the regularity difference is greatest for the high frequency word condition. The reason for the absence of a main effect of syllable regularity and the significance of its interaction with conditions is clear. A word is more easily identified if it has a phonologically correct syllable division, while a pseudoword is more easily identified as such if it has a phonologically incorrect division. Thus, it appears that phonological processing is part of the lexical decision process; both words and pseudowords appear more wordlike when they are regularly syllabified.

For the previous analysis, the basic datum was the average of (up to) three correct latencies per subject for the three stimuli in each combination of time, condition and regularity. Two additional analyses were performed, both of which were concerned with controlling for specific stimulus effects. For the first analysis, stimuli were grouped in pairs such that both members were the same word or pseudoword but differed in regularity, one being the regularly syllabified form and the other the irregularly divided form. For this analysis of variance of latencies, a response was included in the analysis only if responses to both members of a pair were correct. Thus, if certain stimuli were more easily judged by subjects than other stimuli (because of differences in orthography or for other reasons), they would not contribute unsystematically to the regularity effect or its interactions. Table A in the Appendix contains the mean latencies for each stimulus in each experimental condition. A child contributed to the regular and irregular mean latencies of a stimulus only if he or she was correct on both. The number of subjects contributing to each pair is given in parentheses. The analysis of variance of these latency data produced results quite similar to the original latency analysis, although generally with slightly lower (but still strongly significant) F-ratios. In particular, the conditions by regularity interaction was again significant, F(3,102) = 6.31, $MS_e$ = 9114, $p < .001$.

The second analysis for specific stimulus effects was concerned with the generality of the findings with the present stimuli. The analyses done so far

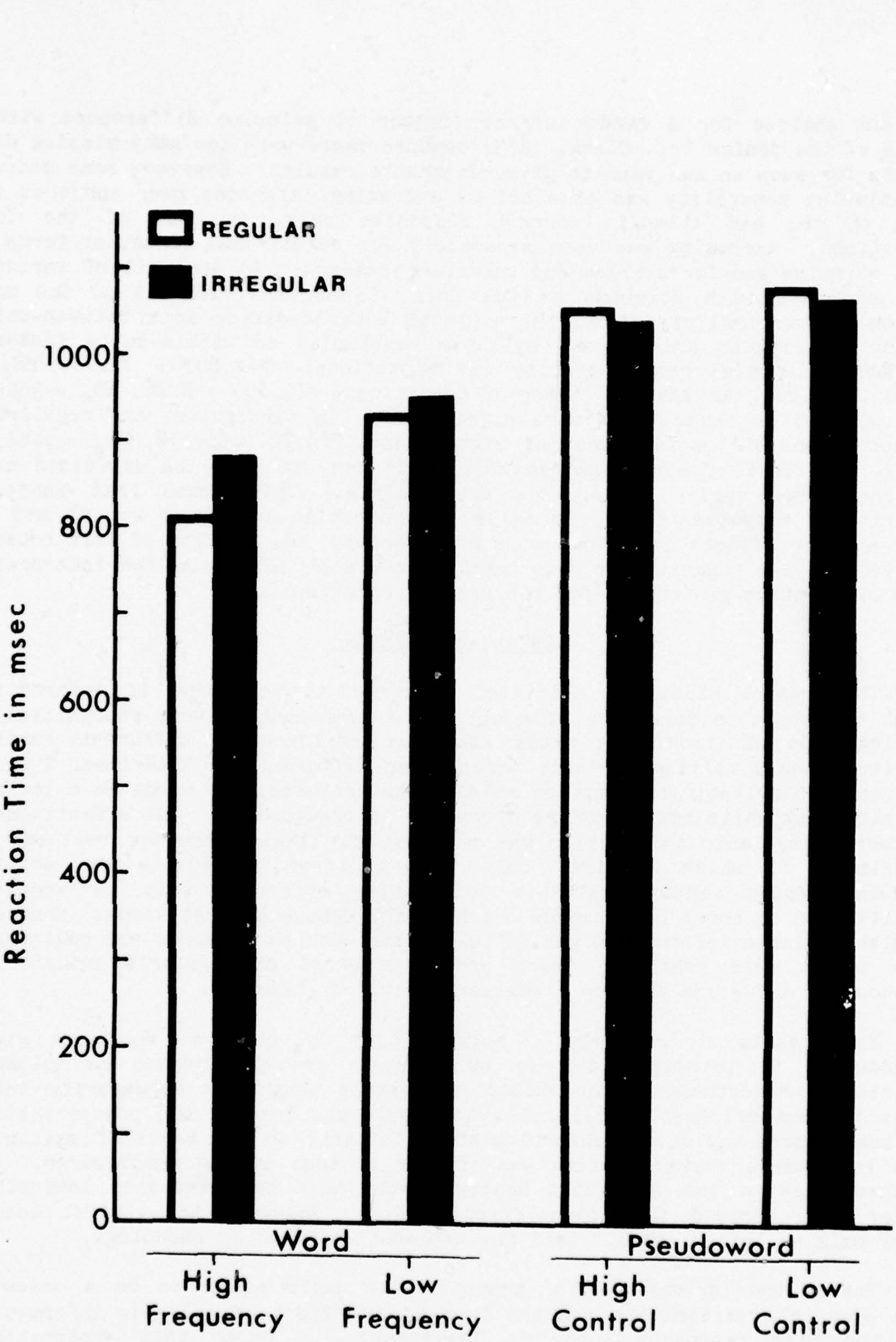


Figure 1. Mean reaction times for regularly and irregularly divided words and pseudowords in each condition (Experiment 2).

did not analyze for a random effects factor of stimulus differences within cells of the design (cf. Clark, 1973) because there were too many missing data points for such an analysis to give reasonable results. However, some measure of stimulus generality was obtained by averaging latencies over subjects for each of the six stimuli (correct responses only) in each of the four conditions. Averaging was done separately for regular and irregular forms of each stimulus and for skilled and unskilled readers. An analysis of variance was performed with stimulus as the unit of analysis (instead of the more common unit of analysis, i.e., the subject), with condition as a between-units factor and reader ability and syllable regularity as within-units factors. For this analysis, reader ability was significant, $F(1,20) = 72.75$, $MS_e = 6979$, $p < .001$, as was the factor of conditions, $F(3,20) = 8.26$, $MS_e = 30841$, $p < .001$ and no other terms were significant. In particular, the regularity by conditions interaction was not significant, $F(3,20) = 1.116$, $MS_e = 9448$, $p < .37$. The lack of more significant effects can, in part, be explained away by the large number of missing data points. This meant that subjects contributed unsystematically to cells of the design and there was no way to take subject effects into account. Nevertheless, the results of this conservative analysis suggest that some caution should be placed on the interpretation of stimulus generality for the present experiment.

## GENERAL DISCUSSION

The present study was initiated in order to determine if information about the syllabic structure of a word is a component of word recognition in reading. In addition, the study asked if children of different reading abilities would utilize syllabic information differently. Experiment 2 found that correct syllabic information speeded the decisions for words in a lexical decision task while retarding the decisions of pseudowords. The effectiveness of correct syllabic information was greatest for high frequency real words. Experiment 1, which required only that children articulate the printed stimulus, found regular syllabic information effective only in speeding vocalization of novel pseudowords but not high frequency real words. However, regular syllabic information was effective for both pseudowords and real words when errors were measured; fewer errors occurred on regularly syllabified pseudowords and words than on irregularly divided items.

The results of Experiment 1 suggest that the children who were given pseudowords to pronounce did so by using a procedure based on syllable divisions. Nevertheless, one should not assume that this necessarily indicates the use of phonological rules to completely specify the pronunciation. The pseudowords may have been articulated, in part, on the basis of analogies with real words whose spelling was similar to that of the pseudowords. The pronunciation of the real word analogs could have been accessed lexically, rather than derived from phonological rules. However, the lexical access hypothesis is less plausible than the explanation based on phonology.

Use of the lexical mode of pronunciation would appear to be a somewhat more plausible explanation for the lack of an effect of syllabic information for the high frequency words in Experiment 1. Under this explanation, children who received high frequency words would have accessed the words' pronunciations lexically without first coding syllabic information. That is, lexical access would have been achieved through a direct visual route without a prior phonological encoding. However, this explanation contrasts with both the error data of Experiment 1 and the outcome of Experiment 2, which required lexical access yet demonstrated the presence of syllabic coding on latencies. Moreover, it would be difficult to explain why a lexical mode would be the preferred mode when the response required was overt articulation (Experiment 1), while a phonological mode would be preferred when the response was a covert semantic one (Experiment 2); if anything, the reverse would be expected. Or one might expect no syllable effect at all in either the naming or lexical decision tasks, based on the negative findings of Fredriksen and Kroll (1976), Forster and Chambers (1973) and others that were treated above in the discussion following Experiment 1. Finally, the use of a phonological code may have been disguised by a discrepancy between the final phonological code and the earlier information that determined the onset of articulation; the advantage of a phonologically correct stimulus (e.g., correctly syllabified) may be lost when the response required is a fast articulation (cf. Davelaar et al., 1978, and the present paper's discussion following Experiment 1).

The clearest evidence that syllabic information is utilized in lexical search comes from Experiment 2. It appears that children used syllable units (or units related to the syllable) while searching for the presence of a stimulus in memory. The syllable code was effective in searching for real words, presumably because addressing based on this code is consonant with the structure of the lexicon. For pseudowords, syllabic coding was counterproductive and slowed correct responses. Because the pseudowords were regular in terms of orthography and phonology, individual syllables of these stimuli would be likely to find matches in the lexicon, biasing a decision incorrectly in favor of "words;" only the combination of the two pseudoword syllables would have no entry in the lexicon. Thus, the decision that the stimulus was not a word would be delayed. I found a related result (Katz, Note 1) in a lexical decision task where the pseudowords varied in their orthographic similarity to English; the more English-like the pseudoword, the longer it took to decide it was not a word.

Why are high frequency words affected more by syllabic information than low frequency words? The explanation depends on the notion that syllables can become functional units in the Laberge-Samuels (1974) and Estes (1977) sense. By a functional unit, it is meant that the letters within a given syllable are associated with a unique code. Although several letters are inputs to the encoder, there is but a single output that does not contain specific letter information. It is this unique syllable code that is used to search memory (along with other possible phonological and nonphonological codes). Assume that high frequency words contain more high frequency syllables than do low frequency words and that high frequency syllables are more likely to be unitized. Therefore, when the unity of a high frequency syllable is disrupted by an incorrect syllable division, the result is more disastrous to the encoding of the syllable than when a low frequency syllable is incorrectly divided. Incorrect syllabification of a low frequency syllable is not as disruptive because it is less likely that there exists such a unitized syllable whose encoding can be disrupted.

Note that the important information in the search process described may be visual in modality; it may be more important that the syllables look the same than sound the same. However, the present experiments make no statement about the modality of the syllabic information used except to point out (as discussed in the Introduction) that phonological codes may have counterparts in visual codes because of the intimate connection between speech codes and visual codes in reading acquisition.

The results of both experiments were clear with respect to reader ability differences. In naming and in lexical search, skilled readers were faster than less skilled readers. In none of the experiments did reader ability interact with syllabification; none of the interactions even approached significance. Therefore, it appears unlikely that the reading skill of the children studied depends on the ability to process syllabic information. Rather, it appears that both skilled and less skilled readers can competently use syllabic information, at least by the fifth grade. Inspection of the protocols of the least able readers supported this suggestion. It may well be the case that younger readers differ in their competency to use syllabic information in word recognition, but the present data suggest that no trace of an earlier ability difference in this area (if it exists) can be found by grade five. Clearly, the present investigations should be extended to younger children to look for early effects; it is possible that early difficulties with syllabic information lead to other problems that persist long after the reader has solved the syllable problem.

In conclusion, evidence for syllabic coding in children's word recognition was found. Whether or not such coding is common in natural reading for fifth grade children was not determined. Although general reader ability differences were found in word recognitionn, these differences were not related to the use of syllabic information. Additional study is needed of younger and poorer readers who may not have developed efficient syllabic processing structures. Also, the nature of the syllabic processing mechanism in reading needs to be explored. In particular, it is important to determine if its structure is represented solely in a visual modality, primarily in a speech modality, or in a more abstract mode prior to exploring the syllabic parsing and retrieval processes themselves. Knowledge of the operating modality of the syllable processor (and other processors of phonological or phonologically derived information) will constrain the plausible mechanisms hypothesized for its operation.

## REFERENCE NOTE

1. Katz, L. *The word frequency effect and orthographic regularity*. Paper presented at the Psychonomic Society Meeting, November 1977.

## REFERENCES

Baddeley, A. D. Working memory and reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *The proceedings of the conference on the processing of visible language*. Eindhoven, 1979.

Carroll, J. B., Davies, P., & Richman, B. *Word frequency book*. New York: American Heritage, 1971.

Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 335-359.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Davelaar, E., Coltheart, M., Besner, D., & Jonasson, J. T. Phonological recoding and lexical access. *Memory & Cognition*, 1978, *6*, 391-402.

Eriksen, C. W., Pollack, M. D., & Montague, W. E. Implicit speech: Mechanism in perceptual encoding? *Journal of Experimental Psychology*, 1970, *84*, 502-597.

Estes, W. K. On the interaction of perception and memory in reading. In D. Laberge and S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Forster, K. I., & Chambers, S. M. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 622-635.

Fredriksen, J. R., & Kroll, J. F. Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, *2*, 361-379.

Henderson, L., Coltheart, M., & Woodhouse, D. Failure to find a syllable effect in number naming. *Memory & Cognition*, 1973, *1*, 304-306.

Kleiman, G. Speech recoding in reading. *Journal of Verbal Learning and Verbal Behavior*, 1975, *14*, 323-339.

LaBerge, D., & Samuels, S. J. Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 1974, *6*, 293-323.

Levy, B. A. Reading: Speech and meaning processes. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 623-638.

Liberman, I. Y., & Shankweiler, D. Speech, the alphabet, and teaching to read. In L. Resnick & P. Weaver (Eds.), *Theory and practice of early reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

Marcel, T., & Patterson, K. Word recognition and production: Reciprocity in clinical and normal studies. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. Functions of graphemic and phonemic codes in visual word recognition. *Memory & Cognition*, 1974, *2*, 309-321.

Spoehr, K. T., & Smith, E. R. The role of syllables in perceptual processing. *Cognitive Psychology*, 1973, *5*, 71-89.

# Appendix

## Table A

Latencies in msec and number of correct responses for skilled and less skilled readers when subject was correct on both the regular and irregular forms of a stimulus. Within each block of four words, the order is high frequency to low frequency and regular to irregular. For pseudowords, the order is high frequency control to low frequency control and regular to irregular.

| | | Skilled | | Less Skilled | | | Skilled | | Less Skilled |
|---|---|---|---|---|---|---|---|---|---|
| wa/ter | (17) | 674 | (18) | 808 | wu/ter | (18) | 954 | (17) | 952 |
| w/ater | | 700 | | 847 | w/uter | | 814 | | 864 |
| pos/ter | (18) | 738 | (18) | 883 | wos/ler | (18) | 827 | (17) | 965 |
| p/oster | | 734 | | 782 | w/osler | | 878 | | 938 |
| wan/ted | (18) | 769 | (15) | 892 | lun/ted | (18) | 888 | (18) | 1025 |
| wa/nted | | 821 | | 1007 | lu/nted | | 851 | | 1087 |
| dus/ted | (18) | 750 | (15) | 1097 | sto/ded | (18) | 1037 | (17) | 1370 |
| du/sted | | 786 | | 1117 | st/oded | | 987 | | 1068 |
| sto/ry | (18) | 757 | (18) | 945 | spo/ry | (16) | 1130 | (13) | 1399 |
| st/ory | | 764 | | 797 | sp/ory | | 1043 | | 1309 |
| hol/ly | (14) | 1050 | (15) | 1004 | hob/ly | (16) | 1116 | (10) | 1269 |
| ho/lly | | 869 | | 929 | ho/bly | | 1109 | | 1251 |
| look/ing | (16) | 764 | (18) | 771 | woak/ing | (16) | 971 | (11) | 1238 |
| lo/oking | | 901 | | 1120 | wo/aking | | 1110 | | 1176 |
| soak/ing | (13) | 832 | (13) | 1093 | boak/ing | (17) | 956 | (11) | 1286 |
| so/aking | | 850 | | 965 | bo/aking | | 975 | | 1203 |
| morn/ing | (17) | 788 | (16) | 954 | mern/ing | (18) | 952 | (18) | 1061 |
| morni/ng | | 892 | | 942 | merni/ng | | 959 | | 1134 |
| jok/ing | (14) | 723 | (15) | 872 | jul/ing | (17) | 941 | (17) | 1075 |
| joki/ng | | 949 | | 1176 | juli/ng | | 942 | | 1161 |
| lit/tle | (18) | 794 | (16) | 944 | lut/tle | (18) | 951 | (17) | 1110 |
| l/ittle | | 793 | | 934 | l/uttle | | 990 | | 1107 |
| nib/ble | (16) | 916 | (14) | 1001 | wim/ble | (13) | 999 | (12) | 1047 |
| n/ibble | | 951 | | 1010 | w/imble | | 991 | | 1009 |

# ACOUSTIC CUES FOR A FRICATIVE-AFFRICATE CONTRAST IN WORD-FINAL POSITION*

Michael F. Dorman,+ Lawrence J. Raphael++ and David Isenberg

Abstract. The experiments reported here were designed to identify some of the acoustic cues to the fricative-affricate contrast in word-final position (as in dish vs. ditch). Listening tests prepared from computer-edited natural speech tokens of dish and ditch reveal that each of the following variables can influence the identification of fricatives and affricates: the temporal and/or spectral characteristics of the vocalic interval, duration of a silent interval, presence or absence of a release burst, rise-time of the fricative noise and the duration of the fricative noise. These results indicate that aspects of qualitatively different acoustic information are integrated over a relatively long period of time when listeners identify fricatives and affricates in word-final position. This outcome suggests that neither a single acoustic property detector nor a single natural category can satisfactorily account for the perception of fricatives and affricates.

## INTRODUCTION

In the several experiments reported here we investigate the acoustic variables that cue the contrast between fricative and affricate in word-final position. Our interest in these variables stems from several sources. One is that our knowledge of the relevant acoustic variables comes from very early unpublished experiments conducted at Haskins Laboratories using the Pattern Playback as a speech synthesizer. Since the Pattern Playback did not have a proper noise source for fricatives,[1] these early studies bear replication and extension. Another source is our interest in the temporal interval over which the cues for a contrast may be distributed (Dorman, Raphael & Liberman, in press; Repp, Liberman, Eccardt & Pesetsky, 1978). Investigation of the fricative-affricate contrast in syllable-final position allows us to determine the perceptual salience of acoustic variables that occur before vocal tract closure for the affricate, the perceptual salience of the closure interval itself and the perceptual salience of aspects of the fricative noise that follows the closure interval.

*Throughout this paper, all occurrences of [ɪ] will appear as [I]. Therefore, e.g., read /dI/ as /dɪ/.

+Also University of Arizona.

++Also Herbert H. Lehman College of the City University of New York.

Acknowledgment: Research supported by Grant HD-01994 from the National Institute of Child Health and Human Development and by NIH Post-doctoral Fellowship NS-05493 awarded to D. Isenberg.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

Inspection of the natural speech utterances dish and ditch in Figure 1 suggests three acoustic properties of the fricative noise that vary between the utterances and thus may serve as acoustic cues (Bailey & Summerfield, 1978). These are the release burst, the rise-time of the fricative noise and the duration of the fricative noise. Indeed, Gerstman (1957) has shown that, in absolute initial position, rapid rise-times and brief durations of fricative noise lead to the perception of affricates while slow rise-times and longer fricative noise durations lead to the perception of fricatives. The role of the release burst that can be seen in the spectrogram of ditch preceding the fricative noise has not been previously investigated. In Figure 1 it is also evident that ditch has a silent closure interval before the onset of the fricative noise that is a consequence of stop articulation. There is no corresponding closure interval in dish. Both Kuipers (1955) and Truby (1955) have shown that long closure intervals lead to the perception of affricates while short intervals lead to the perception of fricatives. However, since both of these studies used the Pattern Playback as a synthesizer, they deserve replication. Finally, in Figure 1 we see that the vocalic portions of the two utterances differ in duration (Isenberg, 1978) as well as in the degree of formant movement preceding the closure. The perceptual salience of these variables has not been previously investigated.

The experiments that follow were designed to assess the effects of each of the acoustic variables described above on the perception of the fricative-affricate contrast. In summary, these variables are: (a) the vocalic portion of the utterance, (b) the duration of the closure interval, (c) the presence of a release burst, (d) the rise-time of the fricative noise amplitude envelope and (e) the duration of the fricative noise.

There is more, however, to our experiments than an investigation of the acoustic cues for a particular phonetic contrast. Our data bear on theories of speech perception that suggest that natural categories (Cutting & Rosner, 1974) or detectors of simple acoustic properties mediate the recognition of phonetic identity. If the outcome of our experiments is that each of the acoustic variables described above can cue the fricative-affricate contrast, then it will be clear that a single natural category or acoustic property detector cannot account for the perception of that contrast.

## EXPERIMENT 1

The purpose of our first experiment was to determine whether two aspects of fricative noise onset--the release burst and the rise-time of the fricative noise--have perceptual salience in the contrast between fricative and affricate in word-final position.

### Method

To assess the effect of the release burst, we digitized and stored in computer memory a male speaker's recording of the utterance, "Put it in the ditch." To create the stimuli for one stimulus condition, we used the ditch stimulus as recorded (with release burst) and inserted intervals of silence between the vocalic portion and the burst. The silent intervals ranged in duration from 0 to 100 msec in 10 msec steps. To create the stimuli for another stimulus condition, we removed the burst from the waveform of the

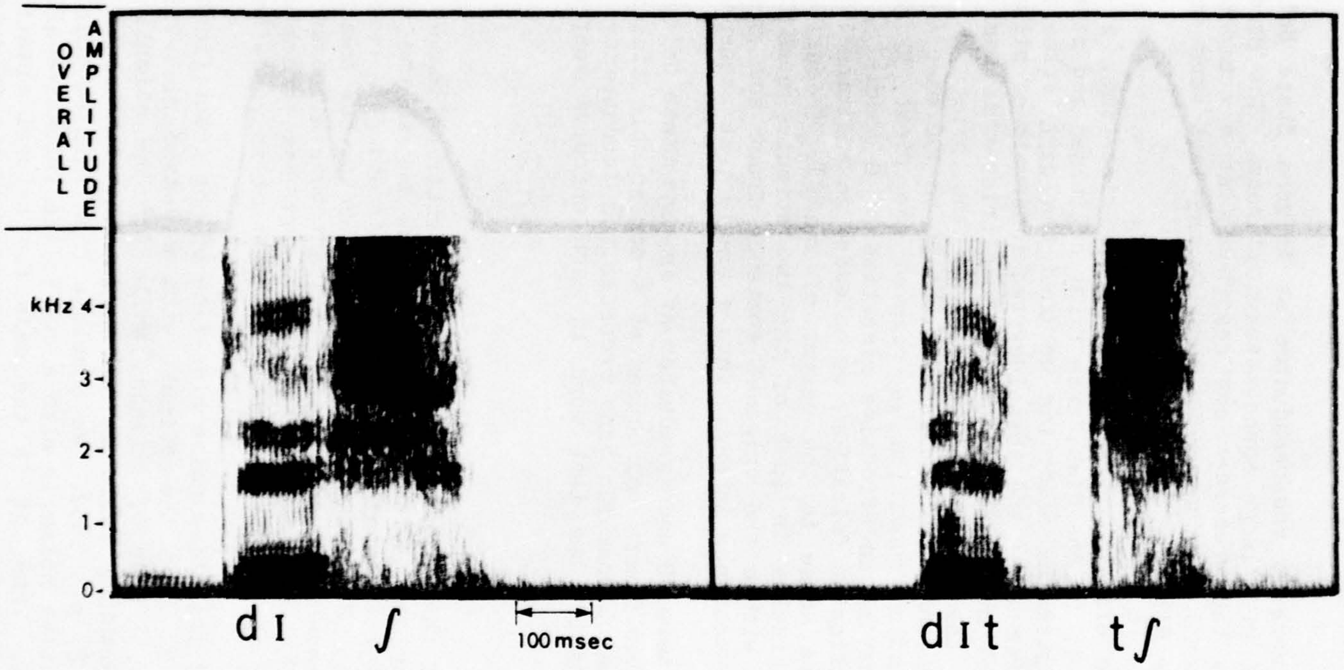


Figure 1: Spectrograms of a male speaker's productions of dish (left) and ditch (right) showing contrastive acoustic features that may be relevant in perception of the fricative-affricate distinction (see text).

ditch stimulus and then inserted the same intervals of silence between vocalic portion and fricative noise onset as were used in the first condition. Four tokens of each stimulus in each condition were generated and the resulting stimuli randomized (with a three-second interstimulus interval) into a single test sequence.

The subjects were 16 undergraduates at Arizona State University. They were run in groups in a large sound-attenuated room. The stimulus sentences were presented via tape recorder and headphones at a comfortable listening level. Subjects were instructed to identify the last word in each stimulus sentence as dish or ditch.

To assess the effects of fricative noise rise-time, a male speaker's recording of "Put it in the dish," was first digitized and stored in computer memory. For one stimulus condition we used the dish stimulus as recorded (with a 35-msec rise-time) and then inserted intervals of silence between the vocalic portion and the fricative noise onset (rise-time was defined as the interval between the onset of energy and the point of maximum signal amplitude). The silent intervals ranged from 20 to 150 msec in 10-msec steps. For a second stimulus condition we removed the first 30 msec of the /ʃ/ fricative noise leaving an effective rise-time of 0 msec.[2] To compensate for the temporal reduction in frication, we embedded an additional 30 msec of full amplitude fricative noise in the center of the /ʃ/ fricative noise. Four tokens of each stimulus in each of the two stimulus conditions were then generated, placed within the original sentence frame and randomized with a three-second inter-sentence interval into a single test sequence.

The subjects were 10 undergraduates at Arizona State University. For all conditions the stimuli were reproduced at a comfortable listening level in a large sound-attenuated room via tape recorder and loudspeaker. Subjects were instructed to identify the last word in each stimulus sentence as dish or ditch.

## Results

The effect of the release burst on the identification of ditch is shown in Figure 2. For the stimuli with a release burst, the phoneme boundary between /ʃ/ and /tʃ/ (the point of 50 percent /ʃ/ and /tʃ/ responses) falls at 18 msec of silence. For the stimuli without a burst, the boundary falls at 28 msec. The difference in boundary is significant by a Wilcoxon sign-ranks test (T = 3, p < .05). We conclude that the burst has perceptual salience in natural speech for the fricative-affricate contrast in word-final position.

The effect of fricative noise rise-time on the identification of ditch is shown in Figure 3. For the stimuli with a 0-msec rise-time, the phoneme boundary falls at 37 msec of silence, while for the stimuli with a 35-msec rise-time, the boundary falls at 57 msec. The difference in boundary is significant (T = 2, p < .05). We conclude from this outcome that the rise-time of the fricative noise is also a cue to the fricative-affricate contrast in natural speech. Thus, it is the case for natural speech, as it is for synthetic speech, that the onset characteristics of the fricative noise play a significant role in the perception of the fricative-affricate contrast.

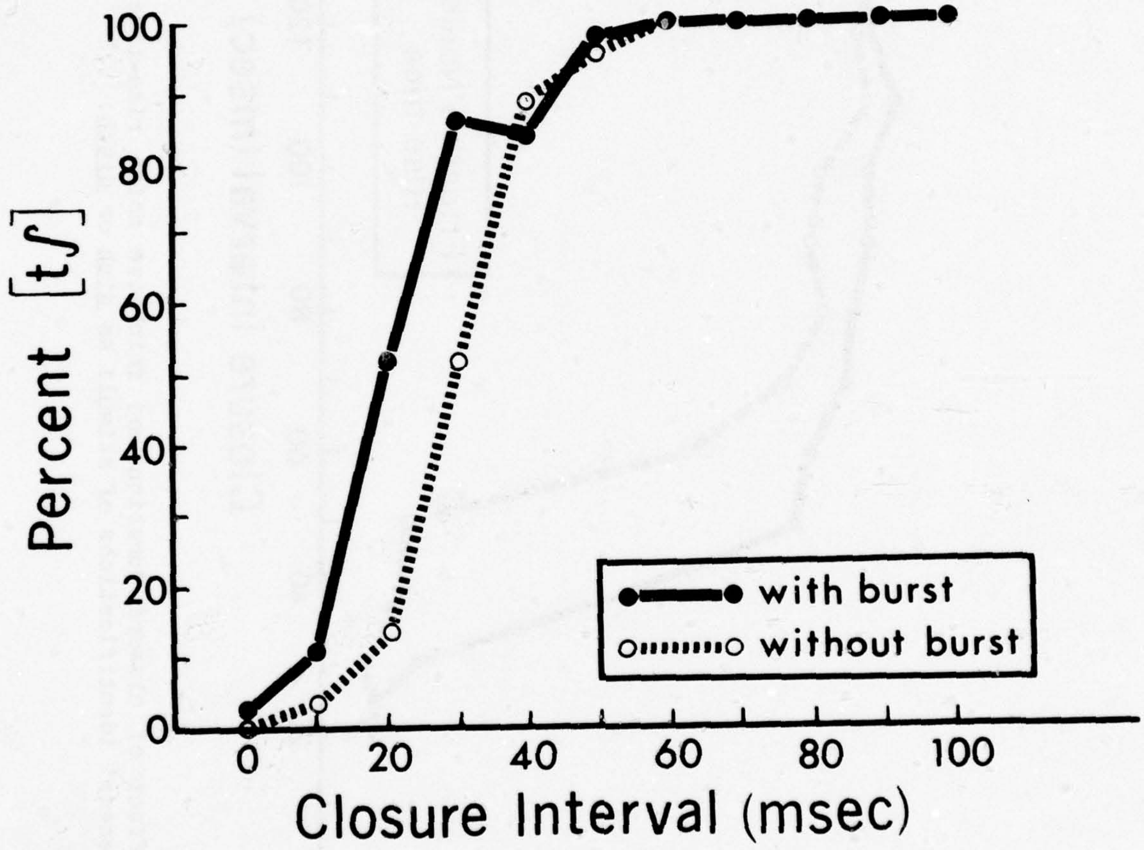


Figure 2: Effect of closure duration and presence or absence of a release burst on listeners' identifications of stimuli as dish or ditch.

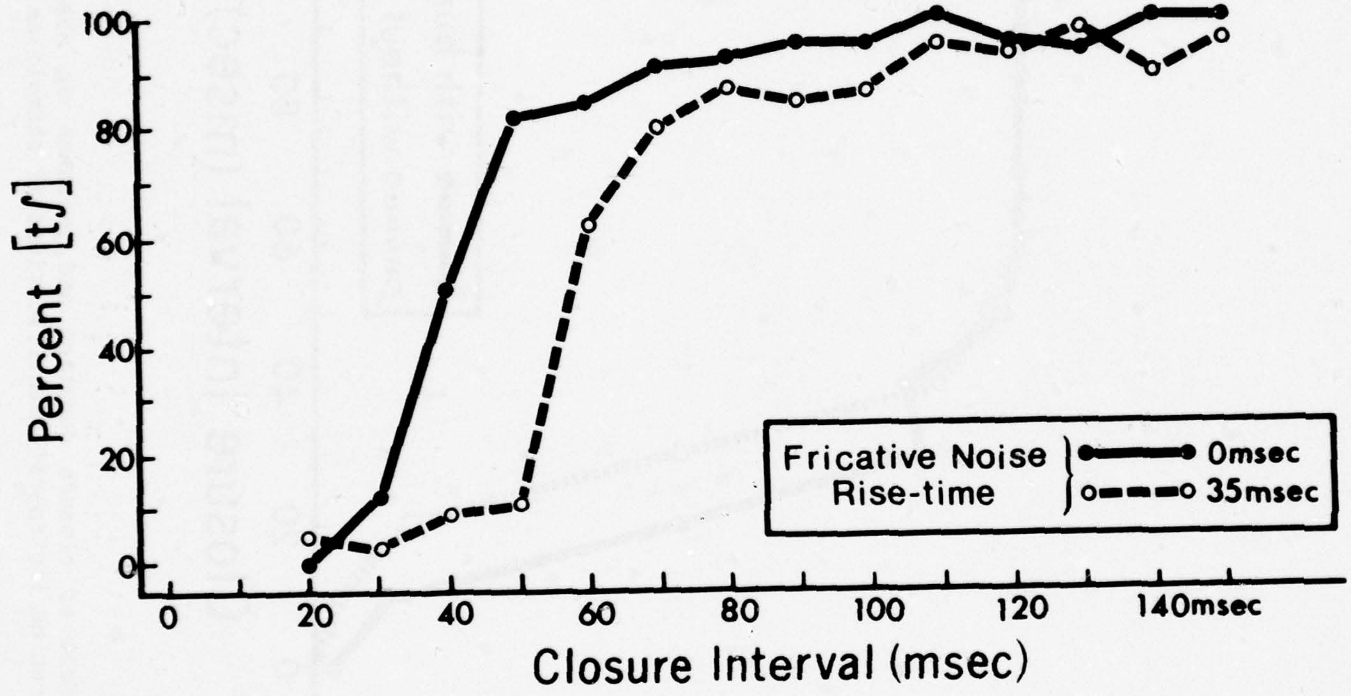


Figure 3: Effect of closure duration and fricative noise rise-time on listeners' identifications of stimuli as dish or ditch.

The duration of the closure interval also exerted a powerful influence on the identification of ditch and dish. Indeed, regardless of whether the source utterance was ditch or dish, at short closure intervals the stimuli were generally heard as dish while at long intervals the stimuli were heard as ditch. Thus, at least two cues, the duration of the closure interval and the fricative noise onset timing, operate to specify the contrast between fricative and affricate in word-final position.

## EXPERIMENT 2

Our first experiment demonstrated that the onset characteristics of the fricative noise (i.e., the rise-time and presence or absence of a burst) and the duration of the preceding silent interval were cues to the fricative-affricate contrast. In Experiment 2 we assessed whether the duration of fricative noise is also a perceptual cue to the contrast in word-final position.

### Method

The utterance, "Put it in the dish," was produced by a male speaker and then stored in digital form in computer memory. To assess the effect of fricative noise duration we created dish stimuli with fricative noise durations of 320, 240 and 160 msec. The stimulus with the longest fricative noise duration was the stimulus originally recorded. To create the stimulus with 240 msec of fricative noise, 80 msec of fricative noise was removed from the center of the 320-msec fricative noise. The onset and offset of the fricative noise were then recombined. A similar procedure with the 240-msec stimulus created the 160-msec stimulus. (We should note that since the fricative noise was aperiodic, the removal and recombination of the fricative noise did not produce audible transients.) For each stimulus condition, intervals of silence ranging from 0 to 100 msec in 10-msec steps were inserted between the vocalic portion and fricative noise portion of the signal. Four tokens of each stimulus in each stimulus condition were generated and then randomized with a three-second interstimulus interval into a single test sequence.

Two groups of 12 undergraduate subjects at Herbert H. Lehman College listened to the stimuli in a large sound-attenuated room. The stimuli were reproduced at a comfortable listening level via tape recorder and loudspeaker. Subjects were asked to identify the last word in each sentence as dish or ditch.

### Results

The effect of the fricative noise duration on the identification of ditch is shown in Figure 4. The phoneme boundary for the 320-msec condition fell at 68 msec of silence, for the 240-msec condition it fell at 61 msec of silence and for the 160-msec condition it fell at 54 msec of silence. The boundaries for the 320-msec and 240-msec conditions were significantly different ($T = 0$, $p < .05$), as were the boundaries for the 240-msec and 160-msec conditions ($T = 5$, $p < .05$). We conclude that the duration of fricative noise is a cue to the fricative-affricate contrast in word-final position.

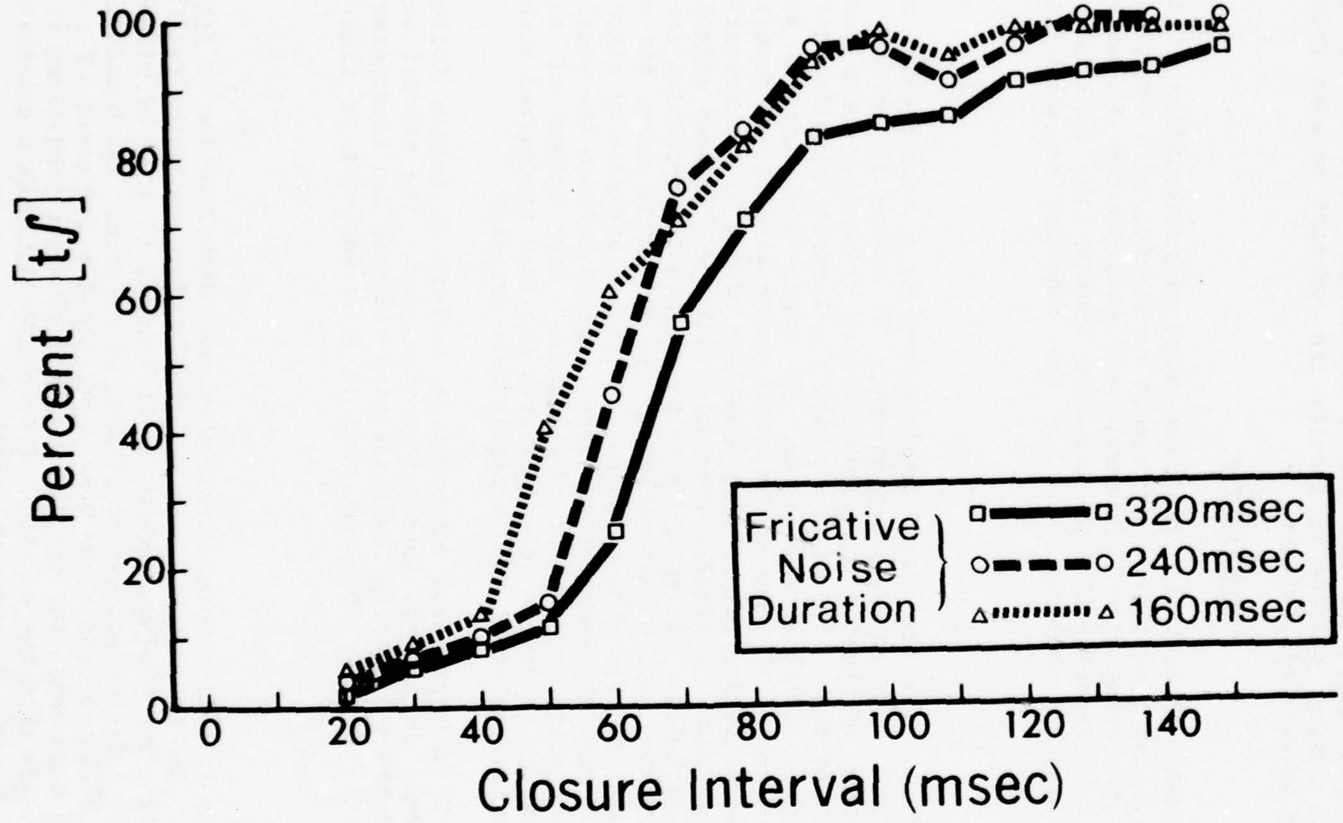


Figure 4: Effect of closure duration and fricative noise duration on listeners' identifications of stimuli as dish or ditch.

In the outcome of Experiment 1 we saw that the duration of the silent interval before the fricative noise and the onset characteristics of the fricative noise contributed to the perception of the fricative-affricate contrast. In the outcome of the present experiment we see that the duration of the fricative noise, as well as its onset, also contributes to that contrast. Thus, we see once again that any single characteristic of the fricative noise is not sufficient to account for perception of a word-final fricative-affricate contrast. Rather, it seems that the perceptual system integrates several types of acoustic information over a relatively long period of time in identification of this contrast (see also Repp et al., 1978).

## EXPERIMENT 3

It is not unexpected that the duration of the closure interval, the release burst, the rise-time of the fricative noise and its duration all affect the perception of fricative-affricate contrasts in word-final position. All vary as a function of the gesture that differentiates fricative from affricate. Extending this reasoning, one might expect that each of the temporally and spectrally distributed acoustic conseguences of that gesture may have perceptual salience. If so, then we should find that properties of the vocalic portions of *dish* and *ditch* also contribute to the perceptual contrast since they differ as a function of whether a fricative or affricate is produced. Indeed, they differ in at least two ways--we know that the vocalic section preceding the affricate is shorter than that preceding the fricative (Isenberg, 1978), and also we see in Figure 1 that the vocalic section produced in *ditch* contains spectral transitions appropriate for complete vocal tract closure that are absent in the vocalic section of *dish*. Experiment 3 was designed to determine whether the vocalic sections of *dish* and *ditch* affect the identification of word-final fricatives and affricates.

### Method

A male speaker's recordings of, "Put it in the dish," and "Put it in the ditch," were digitized and stored in computer memory. To assess the effect of the vocalic portions of *dish* and *ditch* on the perception of the fricative noise, we isolated the vocalic and fricative noise portions of the signals. We then recombined these portions of the signals into four 10-step stimulus series. Each series was defined by the duration of the silent interval, which varied from 0 to 90 msec in 10-msec steps. In order to assess the relative contribution of differences in the vocalic and fricative noise portions of the signal, two of the continua were "crossed." These two continua contained either the vocalic section from *dish* and the fricative noise from *ditch* (/dI/+/tʃ/) or the vocalic section from *ditch* and the fricative noise from *dish* (/dIt/+/ʃ/). The other two continua were "uncrossed" controls. One was constructed from the vocalic and fricative noise portions of *dish* (/dI/+/ʃ/). The other was constructed from the corresponding portions of *ditch* (/dIt/+/tʃ/). All four continua were then randomized into a single test sequence and recorded with a three second inter-sentence interval.

The subjects, 10 volunteers who were personnel at Haskins Laboratories, listened to the stimuli in a sound-attenuated room. The signals were reproduced at a comfortable listening level via tape recorder and headphones.

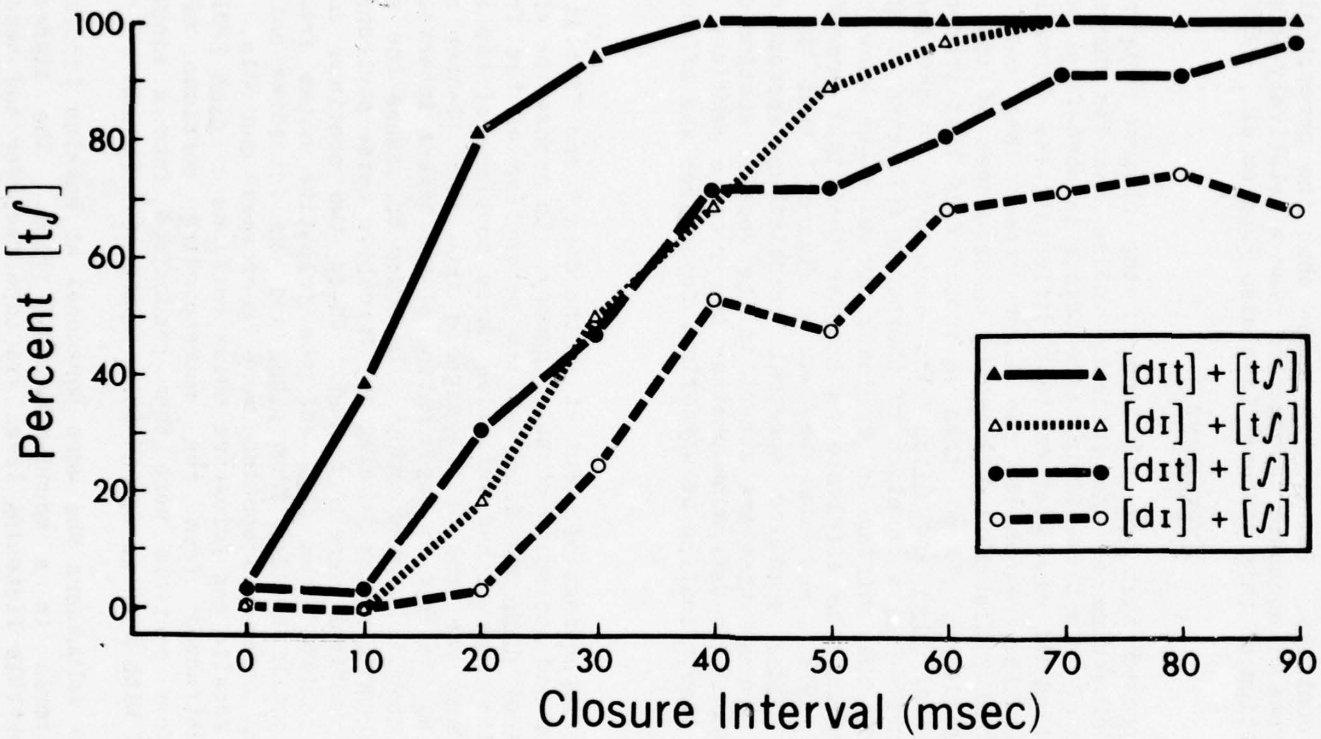


Figure 5: Effect of closure duration and the interaction of vocalic and fricative noise portions of dish and/or ditch on listeners' identifications of stimuli as dish or ditch.

Subjects were again asked to identify the final word in each sentence as dish or ditch.

## Results and Discussion

The identification functions for the four continua are shown in Figure 5. We can see that the phoneme boundary for the /dIt/ + /tʃ/ stimuli falls at the shortest duration of closure interval (11.6 msec). These stimuli also yield the greatest proportion of ditch responses (81.9%). We also see that the phoneme boundary for the /dI/ + /ʃ/ stimuli falls at the longest duration of closure interval (58.9 msec). These stimuli yield the smallest proportion of ditch responses (33.1%). The identification functions for the two "crossed" continua fall between these values, indicating that both the vocalic section and the fricative noise influence the perceptual contrast. The phoneme boundary for the /dIt/ + /ʃ/ continuum (38.6 msec) is not greatly different from that for the /dI/ + /tʃ/ continuum (29.3 msec), nor are the proportions of responses from the two continua (54.1% for /dIt/ + /ʃ/: 65.0% for /dI/ + /tʃ/). In fact, it is clear that the vocalic section has almost as great an effect in specifying this fricative-affricate contrast as does the fricative noise itself.

An analysis of variance was performed on the crossovers computed for each subject in which vocalic section and fricative noise were within-subjects factors. A significant effect was obtained for both the fricative noise [$F(1,9) = 21.96$, $p < .01$] and for the vocalic portion [$F(1,9) = 38.22$, $p < .01$], indicating that both sections were effective in specifying this fricative-affricate contrast. There was no interaction between these factors ($F < 1.0$). A similar pattern was obtained in an additional analysis of variance performed upon the proportion of ditch responses.

These data suggest that properties of the vocalic region are effective cues for the perception of an affricate. One might have predicted that the acoustic properties of the fricative noise would have contained the major cues. However, a comparison of the identification functions for /dIt/ + /ʃ/ and /dI/ + /tʃ/ reveals that listeners make about as many affricate responses in the two conditions. Thus, the vocalic portion of these utterances seems to be about as effective as the fricative noise portion as a cue to the fricative-affricate contrast.

We should point out that a fricative was generally reported even when both the vocalic and fricative noise portions of the signals were appropriate for an affricate, if the silent interval was very brief (less than 20 msec). The converse was also the case--when the closure was very long, an affricate was generally reported. (The failure of the /dI/+/ʃ/ function to reach 100% was due to two subjects who were unable to hear the affricate with these stimuli even at very long closure intervals.) This outcome suggests that silence is a very powerful cue for the perception of an affricate in syllable-final position.

We can suggest three possible accounts for this outcome. One possibility is that when the physical gap in the signal is less than 20 msec, the auditory system does not resolve the gap. On this view, the signal arriving at central processing mechanisms would be without an important cue for stop manner. Therefore, no affricate would be reported. This seems unlikely since there is a systematic increase in the percentage of /tʃ/ responses as silent interval is increased from 10 to 20 msec. Another possibility is that the vocalic portion of the utterance (forward) masks the onset of the fricative noise. This could result in the perception of a more gradual (or fricative-like) fricative noise onset. This explanation, however, seems weak on several accounts. The most important is that we would expect little or no forward masking between signals with such different frequency spectra and source characteristics. Finally, a third possibility is that a very brief silent interval is inappropriate for the production of a stop consonant and is, therefore, not an appropriate cue for the perception of stop manner. This, we suggest, is the most likely account.

## GENERAL DISCUSSION

In this series of experiments with natural speech we have found the following acoustic variables to have perceptual salience for the fricative-affricate contrast in syllable-final position: the vocalic portion of the utterance, the presence and duration of the silent closure interval, the release burst, the rise-time of the fricative noise and the duration of the fricative noise. We may account for the perceptual salience of this constellation of acoustic variables by reference to articulation: Each variable is one of the distributed acoustic consequences of the gesture that differentiates fricative from affricate. In this context we should note Bailey and Summerfield's (1978) suggestion that "...it appears possible to demonstrate that some 'cue-value' attends every acoustic detail that distinguishes two different phonetic events" (see also Lisker, 1977). We see in the present results a confirmation of this suggestion.

The importance of this outcome for theories of speech perception is that those theories must not only specify how the various elements of the acoustic signal are detected, but also what events bear on the same phonetic percept. This latter requirement entails looking across some time-varying segment of the speech signal. We wonder then how the recognition routines "know" what the relevant acoustic hallmarks for a given phoneme are, and how they are spread over time in the speech signal. It seems reasonable to suggest that a system that appreciates the several acoustic consequences of articulatory gestures would be most likely to know what to look for and over what period of the signal to look. This is simply to say that we expect in speech perception a link with speech production (see Bailey & Summerfield, 1978, for an excellent discussion of this issue).

Finally, we should note that the kinds of experiments described here--that is, those that demonstrate trading relationships among the several cues to the fricative-affricate contrast--may well prove probative in investigations of the phylogeny of speech perception. For example, we would not expect nonhuman primates to appreciate the diverse acoustic consequences of the production of fricative and affricate and, thus, we would not expect them to show the trading relations in perception found in the experiments reported here (see also Liberman & Pisoni, 1977).

## REFERENCES

Bailey, P., & Summerfield, Q. Some observations on the perception of [s] + stop clusters. Haskins Laboratories Status Report on Speech Research, 1978, SR-53, 25-60.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sounds of silence in phonetic perception. Journal of the Acoustical Society of America, in press.

Cutting, J., & Rosner, B. Categories and boundaries in speech and music. Perception & Psychophysics, 1974, 16, 564-570.

Gerstman, L. Perceptual dimensions for the fricative noise portions of certain speech sounds. Unpublished doctoral dissertation, New York University, 1957.

Isenberg, D. Effect of speaking rate on the relative duration of stop closure and fricative noise. Haskins Laboratories Status Report on Speech Research, 1978, SR-55/56, 63-79.

Kuipers, A. Affricates in intervocalic position. Haskins Laboratories Quarterly Progress Report, 1955, 15, Appendix 6.

Liberman, A., & Pisoni, D. Evidence for a special speech-processing subsystem in the human. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.

Lisker, L. Closure hiatus: Cue to voicing, manner and place of consonant occlusion. Journal of the Acoustical Society of America, 1977, 61, S-48(A).

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of temporal cues for stop, fricative and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

Truby, H. Affricates. Haskins Laboratories Quarterly Progress Report, 1955, 11, 7-8.

## FOOTNOTES

[1] The pattern playback did not have a separate noise source. Fricatives were simulated by random selection of harmonic tone bursts within the bandwidth of the desired fricative noise (F. S. Cooper, personal communication).

[2] The rise-time of the original fricative noise was 35 msec. When the initial 30 msec of the noise was removed, the rise-time of the fricative noise that remained was essentially instantaneous, although the overall amplitude continued to increase for another 5 msec. Therefore we refer to our stimuli as having rise-times of 35 msec and 0 msec, respectively, even though we removed only 30 msec from the onset of the original fricative noise.

# APPREHENDING SPELLING PATTERNS FOR VOWELS: A DEVELOPMENTAL STUDY

Carol A. Fowler,+ Donald Shankweiler++ and Isabelle Y. Liberman++

Abstract: This study investigates the extent to which children and adults are responsive to orthographic regularities in their readings of nonsense syllables that conform to the phonology and spelling conventions of English words. College students and children of the second, third and fourth years of elementary school read a list of nonsense monosyllables in which most common vowel spellings were presented. Their vowel responses were analyzed according to three categories: incorrect assignment of sound to spelling and correct assignments by context-free and context-dependent criteria. At all levels of reading experience, the proportions of responses falling into the two latter categories far exceeded expectations based on chance responding. These results showed that the children were able to take advantage of orthographic regularities when asked to read unfamiliar words, and, moreover, with increasing age and reading experience they were able progressively to delimit the contexts in which the different regularities apply. The implication is that in learning to read, children do not merely add items to a sight vocabulary by rote recognition of unanalyzed word wholes. Instead, they acquire a practical knowledge of spelling patterns that can readily be applied to words new to them.

## INTRODUCTION

The errors children make in oral reading provide a window through which we may view the special problems of learning to read. One error pattern, in particular, merits further scrutiny because it is found so consistently in the misreadings of the beginning reader of English. We refer to the fact that misreadings of vowels occur with greater frequency than misreadings of consonants; (see Venezky, 1968; Weber, 1970).

In two earlier investigations by our research group (Shankweiler & Liberman, 1972; Fowler, Liberman & Shankweiler, 1977), we reported that vowel misreadings occur about twice as often as consonant misreadings in children's oral reading of isolated words. In the latter paper we considered the possibility that the greater frequency of vowel errors might be an artifact of the syllable structure. Since all the test words used by Shankweiler and

+Also Dartmouth College, Hanover, New Hampshire.
++Also University of Connecticut, Storrs.

Acknowledgment: This work was supported by NICHD Grant HD-1994 to Haskins Laboratories.

Liberman (1972) were monosyllables of the CVC type, the obtained result in that study could have been due entirely to the medial position of the vowel; that is, an embedded segment might be more difficult to apprehend than an initial or a final segment. That possibility was eliminated by the more recent investigation (Fowler et al., 1977), which showed that vowels in initial, medial and final position are equally difficult for beginning readers, and that they generate more errors than consonants regardless of position.

Given this result, we next asked whether the error pattern in reading reflects the linguistic properties of consonants and vowels. Specifically, we might attribute the differential difficulty of vowels and consonants in reading to properties that distinguish them phonologically. We know from speech research that vowels are typically less clearly defined categorically than consonants in speech production and perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Moreover, they are the more fluid and variable of the two classes of phonetic elements, being more subject to phonetic variation across individual and dialect groups. Finally, vowels and consonants have different functional roles in English phonology. For example, vowels are the foundation on which the syllable is constructed and as such are the carriers of prosodic features, while consonants tend to carry the heavier information load.

In view of these major linguistic differences between consonants and vowels, we should expect the misreadings of these phonetic classes to differ not only in frequency but also in the nature of the error pattern. That is indeed what we found in a recent study (Fowler et al., 1977). In that experiment, misreadings were tabulated according to the number of phonetic features shared between the phonological segment of the target word and that of the word as read. Consonant substitutions were found to bear a close phonetic relationship to the consonants of the target word. On the other hand, vowel misreadings were no more similar phonetically to their target segments than would be expected under the assumption of random assignments of sound to spellings.

A difference between the reading behavior with consonants and vowels was expected, but the particular outcome with vowels was puzzling. Why do the child's misreading of vowels not pattern phonetically?

Three possibilities come to mind. One possibility is that the vowels do, in fact, pattern phonetically, but our analysis did not reveal the pattern because vowels are simply not readily amenable to a feature description. Provided that is not the case (and we have no way of assessing this at present), a second possibility is that the children studied by Fowler et al. (1977) tended to adopt a holistic strategy for reading words. Thus, given an unfamiliar word, the child made a guess that was constrained by the identity of the consonants in the word, but was less constrained by the vowels. (Recall that consonants carry the heavier information load in a word.) Thus the children's readings of consonants were accurate or nearly so, while readings of the vowels tended to be random with respect to their target phonemes.

A third possibility is that the children attempted to "sound out" each unfamiliar word by transforming its several orthographic patterns into their phonetic correlates. In English, the spelling-to-sound relationships among vowels are substantially more complex than those among the consonants in the sense that many more vowel than consonant phonemes may be assigned to a given spelling and, similarly, many more vowel than consonant spellings may correspond to a given phoneme (see Dewey, 1970). This characteristic of the vowel orthography may be due in part to the lesser stability of vowels in speech production and perception, as we have suggested.

If children use this analytic strategy in reading by assigning phonological segments to the orthographic patterns of a word, we may expect their misreadings of vowels to bear a relationship to the target phonemes that can be rationalized on orthographic grounds. That is, a child's misreading of a particular vowel spelling should result in the substitution of a phoneme that is possible for that spelling in the context of other words even though it is not correct in the given word. For example, a child should misread have as [heiv] more frequently than as [hεv] because the phoneme /ei/ typically corresponds to the vowel spelling a-e, while the phoneme /ε/ does not.

We do not wish to propose that children adopt either the holistic or the analytic strategy exclusively when they read an unfamiliar word. However, the focus of our research is on the analytic strategy and we have discussed elsewhere its powerful advantages (Liberman, Shankweiler, Liberman, Fowler & Fischer, 1977; see also Gibson & Levin, 1975). Previously, we have assessed the child's ability to analyze the phonetic structure of a written word (Shankweiler & Liberman, 1972; Fowler, Liberman & Shankweiler, 1977). Here we expand that line of research to consider the effect of the orthography on the child's reading behavior. Thus, the present experiment is designed to ask to what extent the children take the orthographic pattern into account as they attempt to read unfamiliar letter strings. Its broad purpose was to obtain evidence that bears on the question of whether with age and experience children learn the regularities of English orthography so that they can generalize to novel instances, or whether they more typically acquire a reading vocabulary in rote fashion.

Venezky and Johnson (Venezky, 1974; Venezky & Johnson, 1973) have provided some evidence related to this question. They have examined the child's practical knowledge of a small number of spelling-to-sound regularities (in particular, the pronunciation of "c" and "g" before e,i,y and a,o,u; the silent-e rule). In general they have found a regular growth with reading experience in the child's ability to assign appropriate sounds to these spelling patterns when they appear in unfamiliar contexts.

Since our previous research has indicated that vowels as a class are difficult for the beginner, we adopt here a somewhat different focus from Venezky's on individual spelling patterns, and investigate a broad and representative sample of vowel spelling-to-sound correspondences.

## EXPERIMENT

The experimental task involves a list of nonsense syllables to be read aloud. Most spellings of English vowels are represented, all with equal frequency insofar as possible. By requiring the child to read nonsense syllables instead of real words, we may obtain a measure of ability to recognize orthographic regularities that is uninflated by rote recognition of familiar words as unanalyzed wholes.

## Method

Stimulus Materials. The nonsense list was composed of monosyllables in which each of 34 English vowel spellings occurred (where possible) in initial, medial and final syllable position. (The spellings are: a, a-e, ai, au, aw, ay, e, e-e, ea, ee, ei, eigh, eu, ew, ey, i, i-e, ie, igh, o, o-e, oa, oe, oi, ou, ow, oy, u, u-e, ue, ui, uy, y-e, ye.) There were 96 items in the list. To equalize, as far as possible, the difficulty of the consonantal context across the different vowel representations, the consonant set included only the stops (b,d,g,p,t,k). Examples of nonsense words in the list are: ud,deg,tuy. The words were printed in lower case on separate unlined 3 x 5 inch file cards. The order of words in the list was random, and every subject received the same random ordering.

Subjects. The subjects were second, third and fourth graders from an elementary school in Andover, Connecticut, and a group of 20 undergraduate students at the University of Connecticut. The names of the elementary school children were chosen alphabetically from the lists of male and female students of each grade. Ten boys and ten girls were tested at each grade level. Testing was done in late fall and early winter.

Procedure. The set of syllables was presented individually to the school children in a single 20-minute session in which (excluding the adult subjects) they also read two real word lists described elsewhere (Fowler et al., 1977). The order of list presentation was balanced across subjects. The adult subjects received only the nonsense list.

The index cards were placed face down in front of the subject and were turned over one by one. The subjects' task was to read each item as it was presented, giving their best guess if they were uncertain how to pronounce it. The children were informed in advance that the items were "pretend" words; the adults were told that the items were "nonsense" words.

The subjects' responses were transcribed by broad phonetic transcription, using the IPA system. An acoustic record was also made on magnetic tape.

Scoring Procedure. Working from the phonetic transcriptions, we scored the responses in two ways: First we considered the vowel produced in response to each test item and asked whether it was a possible reading of the letters representing the vowel, according to the tables constructed by Dewey (1970). Since here we take account of the vowel alone, without regard to its consonantal context, we call it context-free scoring. In the second scoring system we applied conventions[1] of English orthography to the whole syllable and asked whether the vowel as read by the subject was a possible reading of the vowel letters in the particular position in the syllable in which they occurred. Thus, this scoring is context-dependent.

To give an example: the response [teid] to the syllable *tade* is an orthographically possible response according to both context-free and context-dependent scoring systems. It is indeed an instance of a pattern that occurs in many English words, vowel-silent-e, as in *fade* and *made*. However, the response [teid] to the syllable *tad* would not pass the context-dependent criterion, although it would in the context-free system. The response [tid] would fail by both criteria as a reading either of *tade* or *tad*.

## Results

The results of the analysis are presented in Table 1, in which responses are expressed as proportions of opportunity to respond. The "context-free" category refers to the whole set of responses correct by either scoring criterion; the "context-dependent" category refers to the subset of responses that were correct by the more stringent context-dependent scoring system.

TABLE 1: Mean percentage of orthographically possible responses (and standard deviation) in the nonsense list.*

| Grade | Context-free | Context-dependent | Average difference |
|---|---|---|---|
| 2 | 57.7 ( 8.4) | 50.7 (11.5) | 7.0 |
| 3 | 67.0 (12.0) | 60.7 (13.1) | 6.3 |
| 4 | 69.3 (12.5) | 68.0 (11.7) | 1.3 |
| Adult | 81.3 ( 6.3) | 79.3 ( 5.9) | 2.0 |

*See Footnote 2

The most notable result is the extent to which the proportions of correct readings of the vowels exceed the chance value[2] by either the context-free or the context-dependent criterion. Even among second year pupils, the proportion of correct responses exceeds chance by a factor of 2.6. Thus, even readers with just over a year of instruction in reading and writing are able to utilize the ruleful relationship between sound and spelling to decode new items.

The proportion of responses in the context-free and context-dependent categories was subjected to an analysis of variance with one between-groups factor, grade level, and one repeated measure, scoring method. The proportion of correct readings in both categories of scoring method increased with age and experience [F(3,76) = 17.81, p < .001].

The acquisition of orthographic rules is by no means complete in the oldest children we tested (fourth year in the elementary school); adults surpass them by a wide margin ($p = .02$ according to a Scheffé's test). Because the test items were nonsense syllables and not actual words, the increases cannot be attributed to an increase in the size of the vocabulary of familiar words that can be recognized holistically. Apparently readers continue to acquire orthographic rules or intuitions as reading skill improves.

The figures in the third column of Table 1 indicate that not only do children acquire new spelling patterns for vowels as their reading experience grows, but they show increasing sensitivity to the contexts in which each of the possible spellings apply (as in our earlier example, the silent-e marker). Again, even the youngest children (second year) show considerable discretion in their choices from among the possibilities; that is, their performance well exceeds chance by the more stringent context-dependent criterion as well as by the context-free criterion.

Turning to the fourth column of Table 1, we may note the average amount by which performance, as assessed by the more lenient criterion, exceeds performance by the more strict criterion, that is, the context-free/context-dependent difference. That difference decreases systematically between the second and fourth years as the context-dependent responses come to constitute an increasingly greater proportion of the context-free responses. A progressive increase in the proportion of context-dependent responses can be taken to mean that the reader is acquiring context-sensitive spelling-to-sound regularities as his experience increases, since there is no reason to expect that the proportion of context-dependent responses should change if the reading vocabulary expands merely by rote learning. In the present data, the decrease between the context-free and context-dependent responses is marginally significant [$F(3,76) = 2.30$, $p < .08$]. However, as we will show now, the same pattern appears in an analysis of data on the reading of real words obtained in a previous experiment (Fowler et al., 1977).

The real-word data were collected from the same elementary-school children as the nonsense data. Briefly, the real word list included 63 monosyllables in which seven vowel phonemes, /i/,/e/,/a/,/ai/,/au/,/ɔ/, and /ou/ appeared three times each in the initial, medial and final positions in the monosyllable. All were words selected to be familiar by sound to second grade children. Additional information concerning the properties of the real word list is given in Fowler et al. (1977).

The analysis performed for the present purposes on the real word data is analogous to that performed on the nonsense data. The results of the analysis are shown in Table 2. The higher overall performance level on the real word lists as compared to the nonsense list is consistent with previous findings (Liberman, Shankweiler, Orlando, Harris & Bell-Berti, 1971). Familiarity of the items of the real-word list would probably account for the difference. The results of the analysis are in good agreement with the findings on the nonsense words with regard to the point at issue. As in Table 1, the context-free/context-dependent difference decreases with the age and experience of the reader. In these data, the difference is significant [$F(2,57) = 8.85$, $p < .01$]. Thus we may be fairly confident that this decrease is reliable.

TABLE 2: Mean percentage of orthographically possible responses (and standard deviation) in the real word list.*

| Grade | Context-free | Context-dependent | Average difference |
|---|---|---|---|
| 2 | 73.7 (13.6) | 63.3 (19.5) | 10.4 |
| 3 | 84.3 (15.3) | 78.0 (16.9) | 6.3 |
| 4 | 90.0 ( 7.7) | 86.3 (11.3) | 3.7 |

*See Footnote 2

A final analysis was performed to assess the degree of relationship between a child's ability to read real words and his ability to apply spelling-to-sound rules. The analysis was intended to provide an indication, albeit indirect, of the contribution of spelling-to-sound decoding skills to skill in reading individual words.

For the purposes of this analysis, we correlated the percentages of context-free and context-dependent, orthographically possible responses to the nonsense words with a measure of each child's ability to read real words. The second measure was the number of words read correctly among the 63 words of the real word list described earlier. Correlations were computed separately for each grade. The six correlations (three grades by two categories of orthographically possible responses) ranged between .77 and .91 and were all highly significant. Thus, between 59% and 82% of the variation among scores on the real word list is accounted for by individual differences at each grade level in ability to apply spelling-to-sound correspondences. There is no difference in the magnitude of the correlation across grades, nor any tendency for either the context-free or the context-dependent responses to correlate more highly with real-word reading skill.

## DISCUSSION

In this experiment, our purpose was to assess the children's practical knowledge of the regular correspondences between sound and spelling in English orthography. The results show that children know and use these regularities when they are asked to read unfamiliar words. The data, in fact, indicate that readers with as little as one year of reading training[3] rely heavily on their knowledge of orthographic regularities to read unfamiliar words. However, this knowledge is limited in two ways. First, the children's responses are orthographically correct less than sixty percent of the time, indicating that they have not yet acquired the full repertoire of spelling patterns for each vowel spelling. Second, they are less able than the more skilled readers to use context to further restrict the range of possible alternatives. Their knowledge of spelling-to-sound regularities is less context-sensitive than that of more experienced readers. This limitation would be expected to contribute heavily to errors on vowels but much less to consonants (that map in a more nearly one-to-one fashion).

Indeed, our reanalysis of the real-word data reported in Fowler et al. (1977) indicates that the responses on vowels, although they did not pattern phonetically in that experiment, can be characterized as moderately sophisticated guesses based on knowledge of orthographic regularities. That is, when a child made an error assigning a sound to a vowel spelling, he was likely to substitute a sound that is another possible sound for that spelling.

The next experimental question to ask, perhaps, concerns what it is that children learn as their practical knowledge of spelling-to-sound correspondences grows. The present experiment does not help to answer this question, but two possibilities come to mind. On the one hand, children may acquire a set of correspondence *rules* either tacitly or explicitly. Experience in reading, then, serves to add new rules, to add contextual detail to old rules, and to isolate exceptions to rules. Alternatively, a child's knowledge of spelling-to-sound correspondences may consist not of rules, but of a set of probabilities that particular sounds correspond to particular spellings. Thus a child may learn, for example, that *a-e* is pronounced /ei/ two-thirds of the time, /æ/ one-twentieth of the time, and so on.[4] With reading experience, these probabilities may come more nearly to approximate the true probabilities of the English orthography, and the contexts in which they apply may be defined increasingly narrowly.

The rule-governed strategy would be the more efficient of the two for reading unfamiliar words. Thus, if children consistently pronounce *a-e* as /ei/ in unfamiliar words because they "know" that silent *e* makes a vowel "say its name," they will guess correctly two-thirds of the time. But if they guess /ei/ two-thirds of the time in accordance with its relative frequency, they will tend to be correct just four-ninths [$(2/3)^2$] of the time. The better strategy then is to apply the general rule.

Nonetheless, it is not implausible that children may use something like the second strategy. The degree of uncertainty, and thus the consistency with which a child assigns a sound to a spelling, might reflect the frequency with which instances of that particular spelling-to-sound pattern are met in the child's lexicon. In that case, choices would be statistical in nature rather than rule-governed.

In principle, these alternatives may be distinguished by looking at the distribution of children's responses to a particular spelling pattern across many encounters with it when embedded in nonsense words. If their responses are invariant across encounters, then we may assume that they are applying a rule; if the responses are distributed according to the relative frequencies of the relevant spelling-to-sound correspondences, then we may assume that their choices are made on a statistical basis. In practice, the distinction may not be easy to make; first because we do not know what rules the child is applying, if any, and thus we cannot be certain which nonsense words represent instances to which the rule applies, and second because we do not have access to the child's history of encounters with the various spelling-to-sound correspondences. There is no table comparable to that of Dewey (1970) that tabulates the frequencies of these correspondences for school-aged children, and thus we cannot identify instances of statistical behavior with any certainty.

Whatever it may be that a child is picking up as he acquires orthographic structures, analyses of the present nonsense-syllable data and the previous real-word data reveal an orderly increase both in practical knowledge of spelling-to-sound correspondences and the contexts in which they apply. The first increase continued beyond the fourth grade, indicating that as reading skill develops, recognition of orthographic regularities is progressively strengthened. It is clear that in learning to read, children acquire an abstract system, not merely a growing accumulation of items recognized in rote fashion.

In regard to the second increase, in context sensitivity, the fact that there is any difference at all in the younger readers between the proportions of context-free and context-dependent responses in the present data is a reflection of the complexity of English orthography, and the reader's growing appreciation of it. In learning to read a language with a simpler orthography, such as Serbo-Croatian, in which the ideal of "one sound, one symbol" is more closely realized, the difference between context-free and context-dependent responses would have little practical meaning.

## REFERENCES

Dewey, G. *Relative frequency of English spellings*. New York: Teacher's College Press, 1970.

Fowler, C. A., Liberman, I. Y., & Shankweiler, D. On interpreting the error pattern in beginning reading. *Language and Speech*, 1977, *20*, 162-173.

Gibson, E. J., & Levin, H. *The psychology of reading*. Cambridge: M.I.T. Press, 1975.

Hockett, C. F. *Analysis of English spelling: A basic research program*. Cooperative Research Project #639, Cornell University. Ithaca, N.Y.: Cornell University Press, 1963.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, F. W. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. Scarborough (Eds.), *Toward a psychology of reading: The proceedings of the CUNY conferences*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Liberman, I. Y., Shankweiler, D., Orlando, C., Harris, K. S., & Bell-Berti, F. Letter confusions and reversals by sequence in the beginning reader. Implications for Orton's theory of developmental dyslexia. *Cortex*, 1971, *7*, 127-142.

Shankweiler, D., & Liberman, I. Y. Misreading: A search for causes. In J. F. Kavanagh & I. G. Mattingly (Eds), *Language by ear and by eye: The relationships between speech and reading*. Cambridge: M.I.T. Press, 1972.

Venezky, R. In J. Kavanagh (Ed.), *Communicating by language: The reading process*. Bethesda, Md.: U.S. Department of Health, Education and Welfare, p. 203, 1968.

Venezky, R. Theoretical and experimental bases for teaching reading. In T. A. Sebeok (Ed.), *Current trends in linguistics*, Vol. 12. The Hague: Mouton, 1974.

Venezky, R, & Johnson, D. D. Development of two letter-sound patterns in grades one through three. *Journal of Educational Psychology*, 1973, *64*, 109-115.

Weber, R. A linguistic analysis of first-grade reading errors. Reading Research Quarterly, 1970, 5, 427-451.

## FOOTNOTES

[1] The context-dependent subset was difficult to establish. Dewey's (1970) tables are not appropriate for establishing the subset because they do not separately tabulate monosyllables and multisyllabic words. For this reason Hockett's (1963) tables were used here. These tables list all possible phonemic realizations of a spelling as a function of its location in a monosyllable. For the purposes of the present analysis, all sounds listed in Hockett as appropriate for a spelling in the initial, medial or final syllable positions were counted as appropriate responses for that context, provided they were judged to be not too rare.

[2] Chance is approximately 22 percent for the context-free category and 11 percent for the context-dependent category. Chance for the context-free category was computed by summing the possible phonetic realizations for each spelling used in the nonsense-word list (from Dewey, 1970) and dividing by the number of spellings in the list. The outcome of this computation is the average number of possible phonetic realizations per spelling. That value, divided by fifteen, the number of vowel phonemes in English, gives the probability that a given phonetic response to a spelling if it is selected randomly, will be a possible response for that spelling. Chance for the context-dependent category was computed in the same way except that, as explained in the methods section, the criterion for inclusion of a sound as a possible phonetic realization was more stringent.

[3] The approach to their reading instruction was eclectic. One cannot, in the public schools locally, obtain children who have been exposed only to a phonetic approach, or only to a whole-word approach.

[4] The problem of describing the statistical aspects of the orthography can be approached either from the standpoint of the relative frequencies of types or of tokens.

# "PERCEPTUAL CENTERS" IN SPEECH PRODUCTION AND PERCEPTION

Carol A. Fowler+

Abstract. Morton, Marcus, and Frankish (1976) report that listeners hear acoustically isochronous digit sequences as anisochronous. Moreover, given a chance to adjust intervals in the sequences until they are perceptually isochronous, the listeners introduce systematic deviations from isochrony. The present series of studies investigates these phenomena further. They indicate that when asked to *produce* isochronous sequences, talkers generate precisely the acoustic anisochronies that listeners require in order to *hear* a sequence as isochronous. The acoustic anisochronies that talkers produce are expected if talkers initiate the articulation of successive items in the sequence at temporally equidistant intervals. Items whose initial consonants differ in respect to manner class will have acoustic consequences (other than silence) at different lags with respect to their articulatory onsets, thereby generating the observed acoustic anisochronies. The findings suggest that listeners judge isochrony based on acoustic information about *articulatory* timing rather than on some articulation-free acoustic basis.

Sequences of spoken digits are judged to be unevenly timed when the digits are presented at acoustically regular intervals. Moreover, given an opportunity to adjust the intervals to make them sound isochronous, subjects introduce systematic deviations from acoustic regularity (Morton, Marcus, & Frankish, 1976). The deviations are such that long intervals are interposed between a digit starting with a consonant of long acoustic duration and one starting with a short-duration consonant or a vowel (e.g., six-eight); correspondingly short intervals are inserted between the same digit-types presented in reverse order (e.g., eight-six).

These findings disconfirm a hypothesis about listeners' judgments of rhythmicity in speech that, *a priori*, seems simple and plausible--namely that listeners base rhythmicity judgments on the intervals between the onsets of acoustic energy of successive syllables (or, more likely, perhaps, of successive *stressed* syllables) in the sequence.

Morton et al. coin the term "perceptual center" or "P-center" to reference the locus in a word that must be equidistant, temporally, from corresponding loci in surrounding words in order for the sequence to sound

+Also Dartmouth College.

Acknowledgment: This research was supported by NICHD grant HD-01994 and NSF grant NS 13617 to Haskins Laboratories. I thank Betty Tuller and George Wolford for reviewing earlier versions of the manuscript.

[HASKINS LABORATORIES: Status Report on Speech Research SR-57 (1979)]

isochronous to a perceiver. Thus, the perceptual center is the "psychological moment of occurrence" of a word.

The interval-adjustment technique used by Morton et al. did not enable them to locate the P-centers of the different digits, but only to discover the relative temporal alignments of the acoustic onsets of successive digits in a sequence that would make the sequence sound isochronous. As noted, the critical variable affecting the temporal alignment of a digit with respect to surrounding digits was the duration of its acoustic energy prior to the acoustic onset of its vowel. Morton et al. were unable to discover any obvious acoustic markers of the P-centers. They excluded as markers the onset of the word, the onset of the stressed vowel, and the word's or vowel's peak intensity.

Two other experimental investigations, both designed to discover the locus of the stress beat in a word (Allen, 1972; Rapp, Note 1) apparently do pinpoint the P-center, although neither demonstrates how it is marked acoustically. In Allen's study, subjects listened to sentences, each presented repeatedly on a tape loop. During a block of 50 repetitions of the same sentence, subjects tapped their finger "on the beat" of a designated syllable in the sentence. Over blocks of trials, subjects tapped to different syllables in the sentence, and over experimental sessions they listened to different sentences. Allen found that subject's taps tended to be located near the onset of acoustic energy for the vowel in a stressed syllable, but to precede the vowel's onset by a duration that correlated positively (r = .6) with the duration of acoustic energy of the prevocalic consonant or consonant cluster. For example, the tap preceded the vowel onset by 19 msec on the average when the vowel followed a (short-duration) voiced stop, but by 96 msec when it followed a (long-duration) consonant cluster. Rapp (Note 1) found a much higher correlation of precisely the same kind when she asked talkers to repeat various nonsense words in time with a regularly occurring pulse. In this instance, talkers located a nonsense disyllable on the pulse so that the pulse preceded the (second) stressed vowel's acoustic onset by a variable duration. The duration varied directly with the duration of the prevocalic consonant or consonant cluster.

It is evident in Rapp's data that, despite the progressive shift backward of the stress beat with increases in prevocalic consonant duration, the stress-beat tends also to follow the acoustic syllable onset by increasingly longer intervals the greater the prevocalic consonant duration. Assuming that stress beats are P-centers and that P-centers are aligned temporally in perceptually rhythmic utterances, this latter outcome of Rapp's predicts the acoustic anisochronies reported by Morton et al. That is, syllables with acoustically long prevocalic consonants should be located temporally closer to their predecessor than syllables with acoustically short-duration consonants. Thus, evidently, the data of Rapp and of Allen are compatible with those of Morton et al., but add information about P-center location in a word or syllable.

These three rather different experimental procedures agree that a word's stress beat or psychological moment of occurrence does not correspond to the word's acoustic onset, to the acoustic onset of its stressed vowel or to any other obvious acoustic marker. Two of the studies locate the beat within the prevocalic consonant or consonant cluster, but at a variable distance both from the vowel's acoustic onset and from the onset of the word.

Among other areas of investigation, these findings require consideration in examinations of the production and perception of stress-timed rhythms in English. In respect to production, several linguists have claimed that English is a stress-timed language--a language, that is, in which the intervals between major stresses in a naturally produced utterance are fairly uniform in duration (see, for example, Abercrombie, 1964; Classe, 1939, cited in Lehiste, 1973; Pike, 1945). But these claims are based primarily on intuitive judgments and they may have been disconfirmed by several experimental tests. Tests of the stress-timing claim have shown clearly that the acoustically defined intervals between stressed-syllables vary enormously in duration within an utterance (Lehiste, 1973; Duckworth, Note 2; Lea, Note 3; Shen & Peterson, Note 4).[1]

The work of Morton et al., however, suggests a need to reexamine these disconfirmations. If it is the case that talkers regulate the same intervals that listeners judge--that is, those between P-centers and not those between stressed syllable onsets--these disconfirmations may be spurious. The variability among inter-stress intervals might be substantially reduced were the intervals measured that talkers in fact regulate.[2]

In the perceptual domain, some investigators (for example, Lehiste, 1973; Coleman, Note 6) have suggested that stress-timing is largely an imposition by a listener, not by a talker. Indeed, Lehiste finds that naive listeners (no less than the linguists cited above who first posed the stress-timing claim) perceive speech sequences, but not nonspeech analogues, to be more stress-timed than acoustic measurements substantiate (see also Coleman, Note 6).

However, a proposal that listeners _impose_ a stress-timing rhythm on an arhythmic utterance is difficult to rationalize, and it is rendered implausible by other perceptual data. Some evidence, obtained primarily in phoneme-targeting experiments (e.g., Cutler, 1975; Shields, McHugh, & Martin, 1974; see also Martin, 1970), suggests that subjects know with some precision _when_ a stressed syllable (but not an unstressed syllable) is due to occur in a sequence they are monitoring. This evidence implies strongly that spoken utterances have some stress-based rhythm that listeners are able to track. The simplest stress-based rhythm that the perceptual data promote as plausible is stress timing. The evidence of Morton et al., Allen, and Rapp suggests that the rhythmic intervals may be bounded by P-centers. Thus, it is possible that the acoustic measurements of stress timing in the Lehiste study might better have mirrored listeners' judgments had the interval-edges been established at the P-centers.

The suprasegmental rhythms of speech, including stress-timing perhaps, most probably are symptoms of those aspects of articulatory control having to do with coordinating the various structures of the vocal tract, the larynx and the respiratory system. In particular, they may manifest the workings of articulatory controls whose regulated events are temporally more coarsely-grained than the durations of individual phonetic segments. A proposal that speech is stress-timed, then, is a claim about a talker's style of coarse-grained articulatory control, which suggests that one controlled, coherent event in speaking is the interval between stresses. A proposal like this is fairly easy to understand if the inter-stress interval begins and ends at the onsets of stressed syllables. It is far less comprehensible, on the surface anyway, if the interval is bounded by P-centers, since P-centers do not correspond in any obvious way to the edges of linguistic or acoustic units.

The present experiments were designed to investigate further the perceptual phenomenon found by Morton et al. In particular the experiments had two major aims. The first was to rationalize the finding that listeners require perceptually stress-timed utterances to be acoustically anisochronous in particular ways. Morton et al. sought an acoustic explanation for the phenomenon. Here an articulatory account is developed. The second experimental aim was parasitic on the first; it was to obtain information enabling an interpretation of the stress-timing hypothesis, preparatory to reevaluating it.

The first two experiments ask, respectively, whether talkers produce anisochronous rhythms when they are told to talk rhythmically and, if so, whether their deviations from isochrony are just those that listeners require in order to perceive a sequence as stress-timed. Experiments 1 and 2 provide affirmative answers to both of these questions. The third experiment generalizes the production findings to a slightly more natural type of utterance. The final two experiments further examine an articulatory account of the acoustic deviations from isochrony reported by Morton et al.

The results of these experiments suggest that talkers may well produce stress-timed _gestures_ of the vocal tract on request. However, these gestures of the vocal tract produce an anisochronous acoustic signal if the initial consonants of successive syllables are different (primarily, if they differ in manner class). For their part, listeners behave as if the acoustic speech signal provides information about the timing of the talker's _articulatory_ gestures, and as if they base their rhythmicity judgments on that acoustic information about gestural timing. In this respect, the evidence of the present experiments concerning the perception of articulatory rhythms is highly compatible with that of Liberman and his colleagues (see Liberman & Pisoni, 1977, for a review of these studies) relating to the perception of phonetic segments. In particular, both sets of studies show that silence in an utterance provides critical information to a perceiver of speech, putatively about those articulatory gestures that have silence as an acoustic correlate.

The present experiments provide a simple articulatory account of the acoustic deviations from rhythmicity reported by Morton et al. However, they do not immediately explain the variable locus of the P-center as found by Allen (1972) and Rapp (Note 1). Thus they do not suggest what the boundaries of a stress-timed interval might be. A plausible hypothesis about P-center identity can be formulated in articulatory terms, however; this hypothesis will be given in the General Discussion.

## EXPERIMENT 1

The first study asks whether talkers produce acoustically stress-timed utterances when instructed to, or instead, if their utterances deviate from stress-timing in systematic ways.

**Method**

One subject, naive to the purposes of the experiment, was asked to produce a series of nonsense sentences each composed of 6 monosyllables. The monosyllables all rhymed with /ad/, but differed in initial consonant. A nonsense sentence was either homogeneous in composition--that is, its component words were the same (e.g., "mad mad mad mad mad mad")--or it was alternating. An alternating utterance was composed of two different nonsense words produced in alternation (e.g., "mad sad mad sad mad sad"). The syllables ad, bad, mad, nad, tad, sad, composed the vocabulary from which the homogeneous and alternating utterances were constructed. Six homogeneous and 15 alternating utterances (each syllable paired with every other) were produced in all.

The subject, a male adult, was asked to speak at a slow rhythmic rate, stressing every syllable. His utterances were recorded on tape and sound spectrograms were made of each one. Compatibly with the measures of Morton et al. and with those of the above cited investigators of stress-timing in speech, measurements were made of the intervals between acoustic onsets of the syllables in each utterance. Measurements were made in millimeters (1 mm = 7.5 msec). To avoid contamination of the rhythmicity effects by utterance-initial and -final lengthening (Klatt, 1976; Oller, 1973; Lindblom & Rapp, Note 7) only the intervals between syllables two and three, three and four, and four and five were measured. These inter-stress intervals will be called respectively $ISI_2$, $ISI_3$ and $ISI_4$.

**Results**

The relevant measurements are presented in Table 1. They are absolute mean differences in ISI duration for the homogenous and alternating utterances. Thus, the first column of data in Table 1 presents the mean difference in duration between $ISI_2$ and $ISI_3$ for the homogeneous utterances. The second column presents the mean difference between $ISI_3$ and $ISI_4$; and the third column, the difference between $ISI_2$ and $ISI_4$. Absolute values of the individual difference scores were used in the computations of the mean difference values.

Table 1: Absolute values of durational differences (in msec) among the inter-stress intervals of homogeneous and alternating utterances in Experiment 1.

| | Homogeneous | | | Alternating | | |
|---|---|---|---|---|---|---|
| | $ISI_2$-$ISI_3$ | $ISI_3$-$ISI_4$ | $ISI_2$-$ISI_4$ | $ISI_2$-$ISI_3$ | $ISI_3$-$ISI_4$ | $ISI_2$-$ISI_4$ |
| Mean | 20.5 | 27.5 | 19.7 | 116 | 120 | 19.4 |
| S | 11.9 | 18.8 | 17.8 | 80.2 | 70.1 | 18.4 |

Table 1 verifies that the homogeneous utterances were produced in a near stress-timed rhythm. The mean deviations from isochrony ranged between 19.7 and 27.5 msec. These deviations are rather small given that the average interval duration was 474 msec for the homogeneous utterances. An analysis of variance (with initial-segment identity as the random factor) shows that the three deviation values do not differ significantly, $F(2,10) < 1$.

In contrast to this are the corresponding values for the alternating utterances. Here Table 1 shows that $ISI_2$ and $ISI_3$ have durations that differ by 116 msec on the average. Similarly, $ISI_3$ and $ISI_4$ differ by 120 msec whereas $ISI_2$ and $ISI_4$ differ only by 19.4 msec. Recall that in an alternating utterance (e.g., "mad sad mad sad mad sad"), $ISI_2$ (the interval between the first "sad" and the second "mad") and $ISI_3$ (the interval between the second "mad" and the second "sad") are different in that their initial and final consonants are reversed one with respect to the other. The same is true for $ISI_3$ and $ISI_4$. But $ISI_2$ and $ISI_4$ are alike--they start with the same consonant (/s/ in the present example) and end with the same consonant (here /m/).

An analysis of variance of the three mean difference scores for the alternating utterances verifies that the values differ significantly, $F(2,28) = 19.6$, $p < .0001$. Scheffe's tests attribute the significant F value to the difference between the value 19.4 and the other two values, 116 and 120. The latter two values do not differ significantly.

Parallel to the perceptual findings of Morton et al., among the alternating utterances of the present study, the <u>degree</u> of deviation from isochrony of an utterance is closely related to the relative (acoustic) durations of the component syllable's prevocalic acoustic portions. For example, the average deviation from isochrony of syllable onsets (i.e., the average of columns 1-3 in Table 1) was 191 for the utterance "bad sad bad sad...", but was only 44 msec for "bad ad bad ad...". The prevocalic acoustic signal in "bad" is less than 5 msec in duration. This contrasts with 120 msec in "sad" and 0 in "ad." The Pearson product-moment correlation between the average deviation from isochrony and the average difference in duration between the component prevocalic consonants of an alternating utterance is .92, a highly significant value ($p < .001$).

## Discussion

Although the experiment provides a rather limited amount of data, the patterning is clear. When the talker was asked to produce stress-timed utterances, his productions deviated from <u>acoustic</u> isochrony in systematic ways. The deviations are like those created by the listener-subjects of Morton et al. out of acoustically isochronous digit sequences. For both the talker of the present experiment and the listeners described by Morton et al., intervals beginning with acoustically long duration consonants and ending with short-duration consonants were long relative to intervals that are just the reverse of this. This compatibility across the two experiments strongly suggests that in order to hear an utterance as stress-timed, listeners require precisely the deviations from acoustic isochrony that talkers create when they are asked to produce a stress-timed sequence. Experiment 2 establishes the agreement between talkers and listeners using the production data from Experiment 1.

## EXPERIMENT 2

### Method

Stimuli. Twelve utterances were selected from those of the preceding experiment. They were chosen from among the 15 alternating utterances and were the 12 whose deviations from isochrony were largest. Two versions of each utterance were constructed using the PCM system at Haskins Laboratories. One version of each utterance consisted of the middle four syllables of the original six-syllable sentence with the ISIs unaltered. The second version was constructed from the first so that the three ISIs were equal or nearly equal in duration. Isochrony was achieved by electronically splicing silence onto an interval, or less often by deleting silence from intervals of the first versions of each utterance. No portion of an utterance was deleted in which acoustic energy was present (except that the first and sixth syllables were deleted as noted above); only silence was added or removed.

The mean deviations from isochrony among the ISIs for the natural and altered versions of the twelve utterances used in this study are respectively 125 msec and 17 msec. These values verify that the altered versions were substantially more homogeneous in interval duration than were the naturally produced versions.

The mean ISI duration for the altered versions was 473 msec while that for the natural versions was 416 msec. This difference was due to the necessity, for the most part, of adding silence to natural intervals to achieve isochrony, rather than deleting silence. Often, deleting a requisite duration from a long interval was impossible because it would have entailed deleting part of the acoustic signal. However, this difference in mean interval duration is not large and, in any case, is comparable to the (unexplained) difference in the mean duration of homogeneous utterance ISIs (474 msec) and alternating utterance ISIs (418 msec) found in the first experiment. Thus, the ISIs of the altered utterances in the present experiment are within the range of ISIs that are naturally produced in utterances with monosyllabic ISIs. But they may better correspond to the ISIs of naturally produced homogeneous and (like the altered utterances) acoustically isochronous utterances than to alternating, anisochronous utterances.

In the case of one utterance ("bad tad bad tad"), silence was added to all three intervals (in approximately equal amounts), but not so as to alter significantly the degree of anisochrony of the utterance. The difference in duration between the longest and shortest ISI in the natural version was 60 msec; in the constructed version it was 53 msec. However, the mean ISI duration in the original verson was 405 msec, while in the altered version it was 465 msec. This utterance-pair was used as a "catch trial" in a way that is explained below.

On each of 12 trials, the natural version of a sentence was paired with its altered, isochronous counterpart. The natural version occurred in first position of the pair on six trials and the isochronous version on the remaining six trials. The ordering of trials with respect to this manipulation was random. There was a 1500 msec interval between the first and second member of each utterance pair; each pair was repeated once. Four and a half seconds intervened between trials.

Procedure. The subjects' task on hearing each pair was to indicate whether the first or the second version was the more "rhythmic." "Rhythmic" was defined for the purposes of the experiment as equality of intervals between syllable onsets. If a subject judges rhythmicity by comparing the intervals between acoustic syllable onsets, then he should choose the isochronous version on each trial. On the other hand, if he listens for the same intervals that talkers regulate (perhaps the intervals between P-centers), then he should choose the natural version on each trial.

The catch trial provides a check on the basis for the subjects' choices. If, rather than choosing the perceptually more rhythmic versions of a sentence pair as instructed, the subjects choose on the basis of something like perceived naturalness (e.g., because the durations of the inter-syllable pauses in the altered versions may be unnaturally long), then they should choose the natural version on the catch trial as often as they choose it on the other trials. But if they are choosing on the basis of perceived rhythmicity, they should choose the natural version of the catch trial about half the time.

Subjects. Ten students enrolled in the Introductory Psychology course at the University of Connecticut participated in the experiment in exchange for course credit.

## Results

Subjects chose the natural anisochronous version of each sentence pair with far greater than chance frequency. On the 11 trials excluding the catch trial, the natural version was chosen 9.8 times on the average. This value differs significantly from the chance value of 5.5 according to a paired t test, $t(9) = 9.22$, $p < .0001$. Half of the subjects chose the natural version on all 11 trials.

On the average, 9 of the 10 subjects chose the natural version on any given trial. One of the natural utterances was chosen by only 7 subjects; the remaining ones were selected by 8, 9 or all 10 subjects. In contrast, 6 of the 10 subjects chose the natural version of the catch trial. This difference was not due to the natural version of the catch trial utterance having been minimally deviant from acoustic isochrony in the first place, since 10 of the 10 subjects chose the natural version of another utterance ("mad sad mad sad") that was as little deviant from stress-timing as the catch trial utterance "bad tad bad tad." If there was indeed a preference on the subjects' part to choose the natural version of each pair regardless of rhythmicity considerations, it was not of sufficient magnitude to account for the overwhelming tendency, in addition, to choose the acoustically anisochronous version of each pair.

## Discussion

The results of the first experiment suggest that when a talker is asked to produce a rhythmic utterance, he regulates something other than the acoustic onsets of its component syllables. The experiment does not establish what interval is regulated. However, whatever the interval may be, Experiment 2 shows that it is the same interval to which listeners attend when asked to judge the rhythmicity of an utterance.

Since talkers are responsible for the acoustic anisochronies of perceptually stress-timed utterances, it is reasonable to look to the dynamics of articulation for their explanation. However, the articulatory account is only of interest if these anisochronies occur in natural utterances as well as in lists of the sort investigated in the studies of Morton et al. and in Experiment 1. Therefore consideration of an articulation-based rationale for the P-center phenomenon is deferred until the generality of the articulatory findings of Experiment 1 is tested.

## EXPERIMENT 3

Optimally, the generality and naturalness of the findings are tested by examining a large corpus of unconstrained conversation. Under these conditions, if words beginning with acoustically long-duration consonants tend to be "located" temporally closer to preceding words than words starting with short-duration consonants, we could conclude that these timing effects are general to the production of speech and are not peculiar to rhythmic nonsense strings. (These results would not permit any conclusions to be drawn about stress-timing in natural speech, since, as the Discussion will suggest, they may occur for reasons unrelated to the production of speech prosody.)

However, at present, a procedure involving unconstrained conversation is difficult to implement properly. First, the conditions under which the P-center phenomenon may be clearly observed have not been mapped out in any detail. There is some evidence (Marcus, Note 8) that the findings of systematic deviations from acoustic isochrony are somewhat obscured when disyllables are substituted for some of the monosyllables in a string. The reason for this is unknown, but it may have to do with special timing regulations that surround the production of unstressed syllables (see Fowler, Note 5). These timing effects may <u>interact</u> with P-center-related effects and may obscure them. Until they are studied systematically, and until their implications with respect to the P-center concept are understood, it is better to control for them than to let them vary freely as they do in ordinary conversation. Therefore in this preliminary investigation, the utterances produced are restricted to sentences that consist only of monosyllabic stressed syllables.

A second consideration that obviates the use of informal conversation as a data base has to do with the phonetic composition of an ISI. Of particular concern, the acoustic interval between a syllable-final consonant and a following phonetic segment may vary with the degree of articulatory incompatibility between the two segments. It is important that this effect on ISI duration not contribute in a biased way to assessments of temporal alignments of stressed syllables as it is likely to in ordinary conversation. Thus, a preferable procedure is to vary the identity of critical syllable-final consonants and that of syllable-initial segments orthogonally rather than at random. In the present experiment, a more feasible, but somewhat less adequate procedure than this was adopted. Here the critical syllable-final consonant was held <u>constant</u> over variation in the following syllable initial segment. The velar stop /k/ was selected as the critical syllable-final consonant, while the neighboring syllable-initial segments had alveolar or labial places of articulation. The production of these segments should be minimally and equivalently incompatible with the production of the preceding /k/.

## Method

The sentence frame used in the experiment was "Jack likes black -----" (borrowed from Lehiste, 1973) and the words acts, bats, mats, gnats, tacks, facts and sacks were spoken in the frame. The sentences were printed on 3 x 5 inch file cards and were presented to subjects in random order. Following several practice runs to familiarize subjects with the utterances, the test trials began. On each trial, the subject was given a file card and was asked to read the sentence on the card at a slow natural rate. Each sentence was produced three times; the same sentence was never produced twice consecutively. Subjects were tested in a quiet room, and their productions were recorded on audio tape.

Subjects. The subjects were two students in the Introductory Psychology course at the University of Connecticut. They participated in the experiment in exchange for course credit.

Measurements. Spectrograms were made from the recorded utterances and the following intervals were measured:

1. The offset of voicing in "black" to the onset of the test word. (This measurement was made rather than that between /k/-release and test-word onset because it was easier to detect reliably. There is no reason to suppose that this choice of interval should affect the results in any way.)

2. The acoustic onset of the test word to the acoustic onset of its vowel (the acoustic onset of a voiced formant pattern).

## Results

For each subject, the measurements made of a given utterance-type were averaged across its three repetitions. Then correlations were computed between the pair of measurements described in the Methods section (the inter-word interval and the test word's prevocalic acoustic duration). For one subject, the Pearson product-moment correlation is -.96 ($p < .001$); for the other subject, it is -.75 ($p = .02$). Thus, long intervals are interposed between "black" and a short duration syllable-initial consonant, while short intervals are interposed between "black" and a long duration syllable-initial consonant. This is precisely the result reported by Morton et al. and replicated in Experiments 1 and 2.

## Discussion

The results of this experiment and of Experiment 1 may be given a simple articulation-based explanation. The independent variable in each experiment is the identity of certain syllable-initial phonetic segments. The initial consonants vary in voicing, place of articulation and, most importantly, in manner of articulation. The acoustic phenomena observed in the data from the two experiments probably arise in part from differences in the manners of articulation of the consonants and in part from differences in other aspects of their articulatory character including the velocities of their closing gestures. Let us consider how these variables might produce the observed acoustic anisochronies.

Some of the consonants, both in Experiment 1 and in Experiment 3, are stops. Stops are produced, in part, by occluding the vocal tract when the place of articulation is reached. Thus, in the production of /b/ and /p/, the lips are shut temporarily; in /d/ and /t/, the tongue tip occludes the vocal tract at the alveolar ridge; and in /g/ and /k/, the tongue body occludes the vocal tract at the velum. In all cases, the acoustic correlate of the occlusion is silence.

In contrast to the stops are the fricatives (e.g., /s,z,f,v/) and the nasals (/m,n/). In the production of fricatives, the vocal tract is obstructed, but it is not occluded entirely. Thus in producing /s/ and /z/, the tongue tip approaches the alveolar ridge, but does not prevent the passage of air through that location. Likewise, for /f/ and /v/, the air passage is restricted at the lips, but is not closed. Therefore, there is no reason, having to do with vocal tract obstruction at least, for a period of silence to precede the production of a fricative. Similarly for the nasals, /m/ and /n/: Here the oral cavity is occluded; however, the nasal cavity is open and allows the air to escape through the nose.

On these articulatory grounds, other things equal, one would expect a relatively long interval between the offset of a syllable and the onset of acoustic energy for a syllable-initial stop relative to a fricative or nasal. Consequently, ISIs ending with a stop should be longer than those ending with a segment of any other manner class. Experiments 1 and 3 verify this prediction.

In addition to this major effect that the manner class of a consonant may have on an ISI, there is another set of possible production-based influences including the articulatory velocity of a consonant's closing gesture, and, for the stops, the closure interval. These variables are not independent of the linguistic variable, manner class; nor are they entirely redundant. Therefore they are considered separately from manner class here.

The recent data of Kuehn and Moll (1976) indicate that vocal tract closing gestures by the primary articulators for consonants differ in velocity as a function of several variables. Among the critical variables affecting articulatory velocity are the distance that the articulator has to move to achieve closure or near closure (the larger the distance, the faster the movement), and possibly the manner class of the consonant (some fricatives may be produced more slowly than stops).

Other data suggest that the voicing characteristics of the stop consonants also affect articulatory velocity, the closing gestures for the voiceless stops being produced more rapidly than those for the voiced stops (see MacNeilage & Ladefoged, 1976, for a review). (Generally, the voiceless/voiced comparison has been made between stops sharing place of articulation, /p/ being faster than /b/, /t/ than /d/, and /k/ than /g/. However, the data of Kuehn and Moll (1976) indicate little if any difference among the voiceless stops in a variable related to time-to-closure--that is, articulatory velocity with the effect of displacement on velocity subtracted out. Therefore, we will assume here that voiceless stops as a class are produced more rapidly than voiced stops.)

Other things equal, these differences in articulatory velocity should lead to differences in the time after articulatory onset that the different consonant segments have acoustic consequences. Syllable-initial consonants that are produced slowly will have acoustic consequences late relative to rapidly produced segments and therefore will terminate an ISI relatively late. In consequence (other things equal), ISIs ending in voiced stops will be longer than ISIs ending in voiceless stops.

Other things are not entirely equal, however. The closure interval itself is longer in voiceless stops than it is in voiced stops (e.g., Lisker, 1957). This offsets the effect of articulatory velocity on ISI duration. However, data presented in Kozhevnikov and Chistovich (1965) and the acoustic data in Rapp (Note 1) both agree that in some contexts the differences between voiced and voiceless stops in articulatory velocity exceed the reverse differences in closure duration. Thus, the summed effects of the two variables, articulatory velocity of the closing gesture (or, better, time to closure) and closure duration, should lead to an earlier onset of acoustic energy (relative to articulatory onset) for voiceless stops than for voiced stops. Port's data (Note 9) disagree with those of Kozhevnikov and Chistovich, and Rapp. He finds that differences among voiced and voiceless stops in respect to closure duration <u>exceed</u> their differential effects on preceding vowel duration--a measure of articulatory velocity of the closing gesture for the stop. However, his data are on stops that follow stressed /i/ or /I/ and that initiate an unstressed syllable. The data of Rapp and of Kozhevnikov and Chistovich are on stressed syllable-initial stops following stressed (Kozhevnikov and Chistovich) or unstressed (Rapp) /a/. Evidently the variables of articulatory velocity of the closing gesture and of closure interval will affect the time after articulatory onset that acoustic consequences other than silence occur. However, the particular effect that they jointly have may differ depending on variables such as the identity of any preceding vowel and on the stress patterning of the utterance.

Experiments 1 and 3 each offer only a single assessment of the prediction that time-to-closure and closure duration of a segment ending an ISI will affect its duration, since, among the stops, only /b/ and /t/ were included as stimuli. In both experiments, the stops initiated a stressed syllable as in the studies of Rapp, and Kozhevnikov and Chistovich. Therefore the appropriate prediction would be that the pre-acoustic articulatory duration of /b/ should exceed that for /t/. In consequence, ISIs ending in /b/ should exceed those ending in /t/. In Experiment 1, ISIs ending in /b/ averaged 486 msec, while those ending with /t/ averaged 430 msec, $t(11) = 2.35$, $p < .05$. In Experiment 3, the interval between the acoustic onsets of "black" and "bats" averaged 403 and 392 msec for subjects 1 and 2 respectively, while that between "black" and "tacks" averaged 370 and 377 msec. These differences are as predicted, but are probably exaggerated relative to expected differences due to the articulatory variables, time-to-closure and closure duration.

Since the consonant-durations themselves were measured by Morton et al. and, following them, in the present experiments from the onsets of their acoustic energy to vowel onsets, voiced stops are measured to be acoustically shorter than voiceless stops (that is, their voice-onset times are shorter). When consonant-durations (and ISIs) are measured relative to the onsets of acoustic energy for the consonants, even within the stops, the kind of negative correlation observed in Experiment 3 may be seen, but for reasons having to do with articulatory velocity independently of concerns for rhythmicity.

Very little is known about the relative times to closure of other acoustic segments (but see Kuehn & Moll, 1976). It is possible that this variable, together with that of segmental manner class, can account for all of the observed variation in ISI durations in Experiments 1 and 3. This cannot be determined at present, however.

The foregoing discussion makes plausible the following hypotheses about the talkers' performances in Experiments 1 and 3 and the listeners' performances in Experiment 2.

1. When a talker is asked to produce a stress-timed utterance, he acquiesces by initiating the <u>production</u> of stressed syllables at temporally equidistant intervals. For utterances composed of ISIs with nonidentical syllable-initial consonants, the consequence of <u>articulatory</u> isochrony is <u>acoustic</u> anisochrony; the deviations from acoustic isochrony can be predicted on the basis of differences in the manner classes of the consonants or in their respective times to closure (minus closure time for the stops). A talker does not try to compensate for the anisochronies in the acoustic signal that these variables introduce.

2. Comparable differences in the temporal alignments of syllables may be observed in natural speech. Their occurrence here does not necessarily indicate that talkers are working to produce stress-timed utterances under ordinary conditions of speech production. By hypothesis, the anisochronies arise for reasons relating to the articulatory properties of the syllable-initial consonants themselves, and provide no information concerning strategies of suprasegmental speech production.

3. When listeners judge the rhythmicity of an utterance, their judgments are based on information about <u>articulatory</u> timing; they are not based on judgments of the intervals between acoustic syllable onsets, but rather (on a first approximation) on judgments of the intervals between articulatory syllable onsets. This implies that, in the appropriate context, acoustic silence provides information to a perceiver about articulatory activity, and thus about the occurrence of a particular class of phonetic segment--a conclusion already reached on independent grounds by Liberman and his colleagues. (See Liberman & Pisoni, 1977, for a review of some of this evidence.) More precisely, when an intra-phase syllable begins with a stop, its articulatory and <u>perceptual</u> onset may begin at the onset of the silent period or within the silent period that precedes the stop.

Together, the first two hypotheses listed above suggest that the acoustic anisochronies that are observed when talkers intend to produce stress-timing are unrelated to any peculiar strategies for producing the prosodics of speech. That is, they do not arise because the talker *intentionally* causes the onsets of acoustic energy for successive stressed syllables to occur when they do. Instead the anisochronies are a by-product of the talker making articulatory gestures at a stress-timed rate. And they are a by-product due to articulatory properties of individual phonetic segments, and not to properties of articulation having to do with speech prosody.

If these proposals are correct, the acoustic anisochronies should show up even when concerns for stress-timing or other aspects of speech prosody are irrelevant. In fact they should be evident when a talker is trying to produce an *isolated* syllable as rapidly as he can. *Ceteris paribus*, a talker's vocal reaction time measured from the onset of an optical signal to produce a CV syllable to the onset of acoustic energy of the syllable, should be shorter for a syllable starting with a fricative or nasal than for one starting with a stop, and should be shorter for a voiceless stop-vowel syllable than for a voiced stop-vowel syllable. Moreover, these differences in vocal reaction time should correlate with measures of the manner class and articulatory velocity of an initial consonant and with the acoustic deviations from isochrony found in Experiments 1 and 3. Experiment 4 tests these predictions.

## EXPERIMENT 4

### Method

*Stimuli and Procedure*. A list of 20 CV syllables constituted the stimulus set. The vowel in every case was /a/. The consonants were the voiced stops /b,d,g/, the voiceless stops /p,t,k/, the nasals /m,n/, the fricatives /f,v.,s,z,ʃ,θ/, the affricates /č,ǰ/ and the semi-vowels and - consonants /w,r,l,y/. The latter will be called collectively semivowels. Affricates and semivowels were added to provide additional information on the way in which articulatory variables may influence acoustic syllable-onset time and therefore vocal reaction time. Affricates, like stops, require an acoustic silent period during which the vocal tract is occluded. Semivowels are like fricatives and nasals in this respect; they do not occlude the vocal tract or require an acoustic silent period.

The stimuli to be produced by the subjects were presented visually one at a time on a CRT screen, and their sequencing was controlled by a computer program. Each syllable occurred exactly once in each block of twenty trials; ten blocks were presented in all. The syllables were randomized within each block, and each block presented a different random order. Likewise, different subjects received different randomizations of the stimuli.

On each trial, subjects first heard a warning bell; 495 msec later a CV syllable (e.g., "BA") appeared on the viewing screen. The screen was covered with opaque paper except for a slit the width of a single line of print. Subjects were instructed to fixate the slit at a location indicated by an arrow drawn on the paper. The CV syllable appeared in this location on each trial. Intertrial intervals were selected randomly from the values 2, 2.5, 3, 3.5 and 4 seconds, so that trials were not rhythmically sequenced.

Subjects were instructed to read each syllable aloud as quickly as possible after it had appeared on the screen.

Reaction times were obtained in two ways. One way was immediate, but somewhat inaccurate, and served only as feedback for the subject. The second way was substantially more accurate. The first method of obtaining reaction times was via a microphone, a voice relay and a millisecond counter interfaced to the computer. The reaction times obtained in this way were printed on the viewing screen after each trial. Subjects were instructed to maintain their reaction times below 500 msec, and this method of feedback was intended to facilitate their doing so.

Vocal reaction times obtained in this way are inaccurate because of the different energy levels at which different consonants may be produced. A CV syllable, acoustically defined, that starts out low in intensity (e.g., /s/) may not trigger the voice relay until the vowel onset, while one that starts out high in intensity (e.g., /b/) may trigger the relay at once.

To get around this difficulty, reaction times were also obtained by recording the warning bells and the subjects' responses on audio tape. Sound spectrograms were made of each trial and reaction times were measured as the time between the onset of the bell and the onset of the acoustic signal for the syllable minus 495 msec.[3] For each subject, the first two blocks of trials were treated as practice trials and were excluded from the analysis of the data. Also excluded were trials on which misarticulations occurred or reaction times above 600 msec. These averaged 4% across subjects and ranged between 1 and 7%. They will not be discussed further.

Subjects. Subjects were 5 students in the Introductory Psychology course at Dartmouth College. They received course credit for their participation.

## Results

Table 2 gives the mean reaction times and standard deviations for each of the five manner classes of consonant. (/w,r,l,y/ are treated together.) The affricates gave the longest reaction times--followed by the stops. These two manner classes were expected to give the longest reaction times although a difference between them was not expected. An analysis of variance to test the differences among the five means yielded a significant F value, F(4,16) = 6.67, p = .002. However, Scheffes tests showed that the significance was due only to the difference between the affricate class and each of the other four groups and that between the stops and the nasals. The differences among the other classes, including most notably between the stops and the fricatives and between the stops and the semi-vowels, are not significant. Nonetheless, the rank ordering of the differences among the means is in the predicted direction showing that consonants that require an initial period of silence (the affricates and the stops) have longer vocal reaction times than those that do not (nasals, fricatives, semi-vowels and -consonants). A Scheffes test comparing the affricates and stops against the other three manner classes of consonant yields a significant outcome, F(4,16) = 4.97, p < .01.

The results for the voiced and voiceless stops were also as expected based on the variables of time to closure and closure duration. Table 3

Table 2: Mean vocal reaction times (in msec) for the five manner classes of consonant in Experiment 4.

| | stop | affricate | fricative | nasal | semi-vowel |
|---|---|---|---|---|---|
| Mean | 349 | 375 | 337 | 323 | 333 |
| S | 12.6 | 41.2 | 15.7 | 8.7 | 12.9 |

Table 3: Mean vocal reaction times (in msec) for voiced and voiceless stops in Experiment 4.

| | | Voiced | Voiceless |
|---|---|---|---|
| Bilabial | Mean | 364 | 338 |
| | S | 15 | 15 |
| alveolar | Mean | 351 | 338 |
| | S | 21 | 27 |
| velar | Mean | 360 | 344 |
| | S | 13 | 13 |

presents the reaction time means and standard deviations for these two classes of stop each partitioned into the three places of articulation. Across all three places of articulation, the voiced stops gave longer reaction times than the voiceless stops. Moreover, except for the /d/-/t/ comparison, this direction of effect was perfectly consistent across subjects. Thus, all five subjects showed faster reaction times for /p/ than for /b/, and all showed faster times for /k/ than for /g/. Three of the five subjects had faster times on /t/ than on /d/. A two-way repeated measures analysis of variance (voicing by place of articulation) revealed a significant effect of voicing, F(1,4) = 8.86, p < .05, but no effect of place and no interaction.

The duration of the prevocalic acoustic signal of these syllables yields a crude, continuous measure of consonant manner class. This measure correlates negatively (r = -.56, p < .05) with vocal reaction time indicating that long reaction times tend to occur for CV syllables whose consonant phonemes are short in acoustic duration, while short reaction times occur for long duration consonants. (This correlation measure was computed on 16 syllables excluding the semi-vowels and consonants. For these latter syllables, the measure of prevocalic duration was difficult to make reliably.) The correlation parallels the outcomes of Experiments 1 and 3 to the effect that long ISIs precede short duration consonants and short ISIs precede long duration consonants.

When the correlation is restricted to the consonants b,m,n,t,f and s (the phoneme set of Experiment 3), the r value is -.72 (p < .05). This value is somewhat less than the values -.96 and -.75 observed for the subjects of Experiment 3, but it suggests a relationship between vocal reaction time and consonant manner class that is of the same sort as that between ISI and consonant manner class as observed in Experiment 3.

Finally, the latency _differences_ among the possible pairings of b,t,m,n and s (the consonant phonemes of Experiment 1) in the present experiment are highly correlated with the deviations from isochrony of the corresponding alternating utterances of Experiment 1 (r = .86, p < .01).

## Discussion

It seems quite likely in this experiment that the time to _start_ producing a response was fairly constant across the different CV syllables. [Nonetheless, there is a nearly significant correlation of -.44 between vocal reaction time and phoneme frequency as measured by the tables of Dewey (1970).] Therefore, differences in reaction time have primarily to do with any differences in the manner class or other articulatory characteristics of a consonant that affect the time at which the segment has some acoustic consequence other than silence. Since variability due to these factors accounts for 74% of the variance obtained in Experiment 1, it seems likely that these factors are the essential ones that give the P-center phenomenon its character.

In themselves, the observations of Experiments 1, 3 and 4 are not surprising once they have been made. They indicate only that talkers generate systematic temporal alignments of syllables when they talk by adopting very simple and obvious articulatory strategies (i.e., stress-timing in Experiment

1 as instructed, either stess-timing or even more simply, producing a syllable immediately on completing the preceding one in Experiment 3; producing a syllable as quickly as possible in Experiment 4, as instructed). It happens that these strategies have complex, though systematic acoustic consequences.

The findings of these three experiments are of interest primarily in conjunction with the observations of Morton et al. (1976) and their verification in Experiment 2. Together the set of findings indicates that listeners track acoustic information specifying articulatory strategy when they perceive an utterance; they do not treat the acoustic signal as a signal that is independent of its vocal-tract source.

The final experiment in this series reexamines the P-center or stress beat itself, and seeks to establish a relationship between it and the underlying articulation of a syllable.

## EXPERIMENT 5

Acoustically defined the stress beat precedes the onset of a stressed vowel by an amount of time that increases with the duration of the prevocalic consonant. This acoustic description is compatible with the data of Allen (1972), Rapp (Note 1) and Morton et al. (1976), as well as with that of Experiments 1-3.

A simple, but speculative, articulatory proposal that also fits these data is that the stress-beat or P-center is time-locked to (but not necessarily identical to) the onset of articulatory activity for the syllable. (This proposal will be refined in the General Discussion.) On this view, the correlations found by Rapp and Allen, and the acoustic anisochronies reported by Morton et al., occur because different manner classes of consonants tend to have acoustic consequences at variable lags with respect to their articulatory onsets.

These two descriptions of P-center locus, the one a source- or articulation-free acoustic description, and the other articulation-based, are indistinguishable in most utterances. However, they can be distinguished by careful selection of prevocalic consonants. In the following experiment, utterances are selected that allow the two descriptions to be tested separately.

Here, as in Experiment 1, sets of homogeneous and alternating utterances were constructed. Homogeneous utterances consisted of the same CV syllable repeated rhythmically 6 times. Alternating utterances consisted of two CV syllables that were alike in place of articulation, but differed in respect to voicing lead. These syllable-types were repeated in alternation three times each. In both utterance-types the component syllables rhymed with /ad/ and the initial syllables were /b,d,g/ either prevoiced or voiced (no voicing lead). Thus, there were three homogeneous utterances in which no syllable was prevoiced, three in which all were prevoiced, and three alternating utterances in which prevoicing and voicing alternated.

Acoustically, prevoiced consonants differ from their voiced counterparts in having a substantial duration of acoustic energy due to voicing preceding the stop-consonant release. In the voiced stops, the release is preceded by silence.

In contrast to this, articulatorily, the two segments are alike except that the vocal cords vibrate during vocal tract closure in the first case, but not in the second. That is (presumably) the onsets of vocal tract closure for the two classes of stop occur at the same or nearly the same temporal distances from the stop releases. In the prevoiced consonant, closure is accompanied by voicing; in the voiced consonant it is not.

The initial, acoustic, description of the P-center or stress beat suggests that the P-center of a syllable with a prevoiced consonant should precede its vowel onset by a longer duration than the P-center of a syllable with a voiced consonant because the prevoiced CV syllable has the longer prevocalic acoustic duration. More particularly, in Rapp's data, the stress-beat nearly coincided with the release of the voiced stop /d/. If this represents the general case for voiced stops, then the stress beat for voiced /b,d,g/ will coincide, or nearly coincide, with the stop release. That for the prevoiced stops will precede it.

In contrast, if P-centers are time-locked to the articulatory onsets of CV syllables, regardless of the consonants' voicing class, they should be located at the same relative temporal distances from any acoustic markers that the two syllable-types share including the stop-releases.

In summary, then, the acoustic description of the P-center leads to the prediction that intervals in the homogeneous utterances, measured from stop release to stop release, should be isochronous, because the P-centers of their component syllables all have the same locations relative to these acoustic markers (or any others). But alternating utterances should be anisochronous on this measure because the P-center of the voiced stop will coincide with the stop release, while that of the prevoiced stop will precede it. The articulatory description suggests that both sets of utterance-types will be isochronous when intervals between stop-release are measured.

## Method

The nine utterances, six homogeneous and three alternating, were presented in a typed list in random order to two subjects. Both subjects are native speakers of English and both are phoneticans.[4] (Prevoiced stops are not distinct phonemes from the voiced stops in English. Therefore untrained speakers would find it difficult to produce the voicing difference systematically.) The subjects were asked to read through the list twice producing each utterance at a slow stress-timed rate. Their utterances were recorded on audio tape in a soundproof booth.

Sound spectrograms were made of each utterance, and two types of measurements were made on each one:

1. the intervals between the acoustic-syllable onsets of syllables two and three, three and four, and four and five (ISIs 2, 3 and 4 of Experiment 1).

2. the intervals between the stop bursts of syllables two and three, three and four, and four and five.

The first set of intervals should be isochronous only for the six homogeneous utterances if the usual P-center results are obtained. Among the alternating utterances, intervals between a prevoiced and a voiced stop ($ISI_3$) should be longer than those between a voiced and a prevoiced stop ($ISI_2$ and $ISI_4$). The second set of intervals, according to the acoustic description of the P-center, should be isochronous only for the homogeneous utterances. The alternating utterances should be anisochronous in the same way as those of measure 1, although the anisochrony should be less pronounced. According to the articulatory proposal, both utterance-types should be isochronous.

## Results

Table 4 presents the outcome when acoustic syllable-onsets delimit the relevant intervals. On this measure, for both subjects, homogeneous utterances are isochronous and alternating utterances are not. An analysis of variance with ISI and utterance-type (voiced, prevoiced and alternating) as repeated measures factors showed the main effects both to be significant, $F(2,2) = 76.97$, $p = .01$ and $F(2,2) = 21.08$, $p < .05$, respectively, as well as their interaction, $F(4,4) = 47.36$, $p = .003$. Scheffes tests on the individual means attribute the significant interaction to differences between the alternating utterances on the one hand and the prevoiced and voiced utterances on the other. In particular, for voiced and prevoiced homogeneous utterances, the three ISIs do not differ significantly. In contrast, for the alternating utterances, $ISI_3$ is longer than $ISI_2$ and $ISI_4$, $F(2,4) = 135.5$, $p = .008$; $ISI_2$ and $ISI_4$ do not differ one from the other. In addition, $ISI_2$ and $ISI_4$ of the alternating utterances are shorter than their prevoiced and voiced counterparts, $F(5,4) = 37.89$, $p = .003$, while $ISI_3$ in the alternating utterance is longer than its homogeneous counterparts, $F(2,4) = 24.83$, $p = .007$. In short, the alternating utterances deviate from isochrony in the predicted way in that intervals starting with prevoiced stops and ending with voiced stops are long relative to intervals that are the reverse of this.

Table 5 presents the comparable values on the burst-to-burst measure. On this measure, all utterance-sets for both subjects are isochronous. (Although the alternating utterances show a slight cyclicity in ISI duration, it is contrary to the predicted direction and in any event is nonsignificant.) The analysis of variance yielded nonsignificant outcomes on both main effects and on the interaction term.

## Discussion

The experimental outcome supports the articulation-based description of the P-center and fails to support an articulation-free acoustic description. It should be noted, of course, that the experimental design stacked the deck somewhat in favor of the articulatory description. Its prediction was that on the critical burst-to-burst measure, the null hypothesis would fail to be rejected. This is a weak prediction.

Two kinds of consideration support the argument that there are *truly* no differences among levels of the independent variable ISI (rather than that

Table 4: Durations (in msec) of inter-stress intervals of homogeneous and alternating utterances in Experiment 5 measured onset to onset. Each value is the average of 6 tokens.

| | | ISI$_2$ | ISI$_3$ | ISI$_4$ |
|---|---|---|---|---|
| prevoiced | AA | 700 | 704 | 696 |
| | LL | 698 | 686 | 684 |
| voiced | AA | 689 | 676 | 698 |
| | LL | 645 | 654 | 665 |
| alternating | AA | 576 | 797 | 580 |
| | LL | 563 | 731 | 555 |

Table 5: Durations (in msec) of inter-stress intervals of homogeneous and alternating utterances in Experiment 5 measured burst to burst. Each value is the average of 6 tokens.

| | | ISI$_2$ | ISI$_3$ | ISI$_4$ |
|---|---|---|---|---|
| prevoiced | AA | 698 | 692 | 726 |
| | LL | 696 | 670 | 690 |
| voiced | AA | 689 | 676 | 698 |
| | LL | 645 | 654 | 665 |
| alternating | AA | 700 | 690 | 706 |
| | LL | 660 | 639 | 649 |

there are differences to which the experimental design is insensitive).

First, it is clear that the data are not very noisy. The average deviation from isochrony for the homogeneous utterances measured from acoustic syllable-onset to acoustic syllable-onset averages 12 msec (/$ISI_2$-$ISI_3$/ and /$ISI_3$-$ISI_4$/ averaged) for one subject and 9 msec for the other. Moreover, the patterning of these data on this measure is similar to that of Experiment 1 and of the other P-center related studies.

Second on the burst-to-burst measure, the prevoiced to voiced interval does not even tend to be long. Instead, it is slightly shorter than the other intervals, but not reliably so. (This may signify that, like voiceless and voiced stops, prevoiced and voiced stops differ somewhat in time-to-closure.)

We conclude, *tentatively*, therefore, that the P-center is time-locked to the articulatory onset of the prevocalic consonant in a monosyllabic stressed utterance.

## GENERAL DISCUSSION

This final discussion will consider three questions:

1. What is a P-center (and, closely related to this, what does it correspond to in a speech event)?

2. How do listeners track P-centers when they make rhythmicity judgments and on other occasions (as in the phoneme-targeting experiments)?

3. What implications does the P-center phenomenon have, if any, for devising and evaluating a plausible proposal about natural suprasegmental speech rhythms, stress-timing in particular?

### *What is a P-center?*

Curiously, P-centers are phenomena that talkers can regulate, and listeners can track, but that investigators are unable to *see* in any optical representation of an utterance (see Marcus, Note 8).

Were it not for the Allen and Rapp studies, the P-center might be supposed to correspond to the articulatory onset of a stressed syllable. This would have different acoustic correlates depending on what type of gesture is initiated. However, the Rapp and Allen studies place the stress beat very often *within* the acoustic realization of the stressed syllable's prevocalic acoustic signal, and these mutually reinforcing results require an accounting. A speculative account is suggested here.

The P-center may correspond to articulatory activity relating to the stressed *vowel* itself (perhaps to its articulatory onset, or to the attainment of the nearest approximation to its "target" vocal tract shape). Anticipatory coarticulation in CV syllables is very well documented. Due to coarticulation, the articulatory vowel onset would tend to occur *during* the production of a preceding consonant. Gay (1977) reports that the initiation of movement towards $V_2$ in a $V_1CV_2$ utterance coincides with, or occurs just after, the attainment of consonantal closure. The consonants in his study were the voiceless stops. This would locate the P-center no earlier than the onset of the consonant's silent period if it corresponds to the vowel's articulatory onset, or somewhat thereafter if it corresponds to the vowel's nearest approximation to its target. Rapp's figure I-B-6 includes one voiceless stop, t. The stress beat in /atád/ was located somewhat after t's release suggesting that the target view of the P-center locus, suggested above, may be the more accurate of the two proposals.

Interestingly, Carney and Moll (1971) report that for $hV_1CV_2$ nonsense utterances in which the consonant is a fricative rather than a stop, movement toward the second vowel begins *before* the attainment of consonant near-closure for the fricative. This difference between the studies of Gay, and Carney and Moll, is consistent with the evidence (Kuehn & Moll, 1976; see also MacNeilage & Ladefoged, 1976) that in general, fricatives have slower articulatory velocities than stops. If consonants and their following vowels have articulatory *onsets* at some constant temporal delay, one with respect to the other, regardless of the manner class of the consonant, then articulatory reference points for the vowel should appear sooner in a fricative's production than in the more rapid production of a stop. This would place a vowel-related P-center relatively earlier in a fricative than in a stop. It is difficult to say whether this is the case articulatorily. However, it seems generally to be compatible with the acoustic data. In Allen's acoustic measurements, taps to syllables preceded by voiced stops were located closest to the acoustic vowel onset (and thus to the consonant offset). Next were voiceless and voiced fricatives. These data are compatible with the foregoing considerations. However, farthest away of all were the voiceless stops, which should have the highest articulatory velocity going into the closure period. Rapp's data offer only limited relevant information, but it is similar to Allen's. Among the stops and fricatives that she examined, the stress beat is closest to the vowel onset for /d/; it is farther away for /t/ and /s/, which are quite similar one to the other. /t/'s stress beat precedes that for /s/ slightly.

It is also interesting in regard to this vowel-based proposal that in Rapp's data, talkers located the stress beat at a nearly invariant interval *after* the first (unstressed) vowel's acoustic onset in the various /a$C_n$ád/ utterances, even though the beat occurred at a variable locus relative to acoustic markers for the consonants and second, stressed vowel.

## How Do Listeners Track P-centers?

Morton et al. failed to uncover an acoustic marker for the P-center. Of course, this failure does not imply that the P-center has no acoustic marker. But coupled with the present findings, it suggests a need to reconsider the kind of acoustic marker that one can reasonably expect to find. If indeed the P-centers are acoustic correlates of some *abstract* gesture-*type* (for example, the onset of articulatory activity for a vowel, any vowel), then the acoustic correlates may not be invariant in any superficial sense. There is no reason to expect the P-center to correspond, for instance, to an intensity peak or to an abrupt change in fundamental frequency if these are not acoustic correlates of its underlying gesture-type. Instead, the acoustic correlates of the P-center may be that (very large) *class* of acoustic signals that in the appropriate context, *signify* (for example) "the onset of articulatory activity for a vowel." The problem is to explain what about these signals, other than some simple shared acoustic property, endows them with that significance for a listener.

Even if simple acoustic invariants are excluded as plausible markers of P-centers, complex acoustic correlates probably are not. Vowel onsets (and asymptotic attainments of "target" vocal tract shapes) may well be marked by their acoustic coarticulatory influences on consonants. Due to coarticulation, consonants vary acoustically with their segmental context. This means that (among other segments) a post-consonantal vowel's "anticipatory" articulation during the consonant is marked *by* its effect on the acoustic signal for the consonant. Its articulatory onset, then, may be marked by the initiation of its effect on the consonant. Its attainment of a target shape of the vocal tract may be signaled by a stabilization of the vowel's effect on the acoustic signal for the consonant. Possibly this information marks the P-center for a listener.

## *Suprasegmental Speech Rhythms*

A convenient aspect of the proposal that P-centers correspond to articulatory *onsets* is that it enables an intuitive view of speech timing to be preserved--namely that timed events begin and end at the edges of linguistic units. If, instead, P-centers correspond to some within-segment locus--e.g., to vowel-target near-attainments or to some variable locus within a prevocalic consonant--and if intervals between P-centers are regulated in speaking (as they clearly *can* be)--current models of speech production based on linguistic segmental and suprasegmental timing units are disconfirmed. Moreover, devising a new model would entail a major overhaul in our views of speech production. Nonetheless the data do seem to support the articulatory targets view more than the vowel onsets view of P-center locus.

In respect to stress-timing itself, as noted in the introductory section, the P-center findings suggest a need to reevaluate the tests of the claims that speech is stress-timed. One necessary adjustment to the procedures cited earlier (among other adjustments; see Fowler, Note 5) is to measure the intervals between P-centers rather than those between acoustic onsets of stressed syllables. But given the difficulty of locating P-centers precisely, and more importantly, given that the stress-timing proposals are not properly tested by measuring inter-stress intervals and looking for significant differences (see footnote 1), other kinds of tests that avoid ISI measurements should be adopted (e.g., Allen, Note 10; Fowler, Note 5).

## REFERENCE NOTES

1. Rapp, K. A study of syllable timing. *Papers from the Institute of Linguistics, University of Stockholm*, 1971, *8*, 14-19.
2. Duckworth, J. *An inquiry into the validity of the isochronic hypothesis*. Unpublished doctoral dissertation, University of Connecticut, 1965.
3. Lea, W. *Prosodic aids to speech recognition: IV. A general strategy for prosodically guided speech understanding*. (ARPA Report No. PX10791), 1974.

4. Shen, Y., & Peterson, G. Isochronism in English. *Studies in Linguistics, Occasional Papers, University of Buffalo*, 1962, *9*.
5. Fowler, C. *Timing control in speech production*. Bloomington, Indiana: Indiana University Linguistics Club, 1977.
6. Coleman, C. *A study of acoustical and perceptual attributes of isochrony in spoken English*. Unpublished doctoral dissertation, University of Washington, 1974.
7. Lindblom, B., & Rapp, K. Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics, University of Stockholm*, 1973, *21*, 1-59.
8. Marcus, S. *Perceptual centers*. Unpublished fellowship dissertation, Kings College, Cambridge, 1975.
9. Port, R. *The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words*. Bloomington, Indiana: Indiana University Linguistics Club, 1977.
10. Allen, G. The place of rhythm in a theory of language. *UCLA Working Papers*, 1968, *10*, 62-84.

## REFERENCES

Abercrombie, D. Syllable quantity and enclitics in English. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (Eds.), *In honour of Daniel Jones*. London: Longmans, 1964.

Allen, G. The location of rhythmic stress beats in English: An experimental study, I. *Language and Speech*, 1972, *15*, 72-100.

Carney, P., & Moll, K. A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica*, 1971, *23*, 193-202.

Cutler, A. Rhythmic factors in the determination of perceived stress. *Journal of the Acoustical Society of America*, 1975, *57*, S25 (Abstract).

Dewey, G. *Relative frequency of English spellings*. New York: Teachers College Press, 1970.

Gay, T. Articulatory movements in VCV sequences. *Journal of the Acoustical Society of America*, 1977, *62*, 183-193.

Klatt, D. The linguistic uses of segment duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Kozhevnikov, V. A., & Chistovich, L. A. Speech: Articulation and perception. Moscow-Leningrad, 1965. (English translation: J.P.R.S., Washington, D.C., No. JPRS 30543.)

Kuehn, D., & Moll, K. A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 1976, *4*, 303-320.

Lehiste, I. Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America*, 1973, *51*, 2018-2024.

Liberman, A., & Pisoni, D. Evidence for a special speech-perceiving subsystem in the human. In T. Bullock (Ed.), *Recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen, 1977.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 1957, *33*, 42-49.

MacNeilage, P., & Ladefoged, P. The production of speech and language. In E. C. Carterette, & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 7). New York: Academic Press, 1976.

Martin, J. Rhythm-induced judgments of word stress in sentences. *Journal of*

*Verbal Learning and Verbal Behavior*, 1970, *9*, 627-633.

Morton, J., Marcus, S., & Frankish, C. Perceptual centers (P-centers). *Psychological Review*, 1976, *83*, 405-408.

Oller, D. The effects of position in utterance on speech segment-duration in English. *Journal of the Acoustical Society of America*, 1973, *54*, 1235-1246.

Pike, K. *Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

Shields, J., McHugh, A., & Martin, J. Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, 1974, *102*, 250-255.

Trager, G. L., & Smith, H. L. *Outline of English structure*, *Studies in Linguistics*, No. 3. Norman, Oklahoma: Battenburg, 1951.

## FOOTNOTES

[1] For brevity, an inaccuracy has been introduced into the discussion. To my knowledge, in every instance in which a stress-timing proposal has been made, it has taken a very weak form to the effect that the intervals between stressed-syllable onsets *tend* towards isochrony. They are held to deviate from strict isochrony to a degree that varies with the compositional (phonological, syllabic, grammatical) heterogeneity of its component inter-stress intervals. However, although the stress-timing claims have invariably taken this weakened form, the experimental tests cited above are uniformly of the stronger (and easier to test) claim that inter-stress intervals are isochronous. The literature suggests that a fair test of the weak form of the hypothesis yields more hospitable data (see Fowler, Note 5, for a review).

[2] The extant evidence makes it clear that this adjustment in measurement strategy would not reduce the variability in inter-stress interval duration to zero. Some more of the variability in the studies of Duckworth (Note 2) and Shen and Peterson (Note 4) may be put down to an infelicitous assignment of major stresses to syllables. Both adopted the conservative view of Trager and Smith (1951) to the effect that only one major stress may occur between two terminal junctures in a sentence. Nonetheless, substantial variability in interval durations was also reported by Lehiste (1973) and by Lea (Note 3), not all of which would disappear were the intervals between P-centers measured. However, it may be argued that the remaining variability is compatible with the weak version of the stress-timing view (see Fowler, Note 5).

[3] I thank Don Nemcek for making the spectrograms.

[4] I thank Arthur Abramson and Leigh Lisker for serving as subjects in this experiment.

# INFLUENCE OF VOCALIC ENVIRONMENT ON PERCEPTION OF SILENCE IN SPEECH

Bruno H. Repp

**Abstract.** The amount of intervocalic silence needed to perceive two different stop consonants in synthetic stimuli of the type /$V_1$b-g$V_2$/ (single-cluster boundary), and two identical stops in stimuli of the type /$V_1$b-b$V_2$/ (single-geminate boundary), was determined as a function of different vowel contexts ($V_1$, $V_2$ = /i/, /a/, /u/) and of different durations of the initial and final vocalic portions (120, 180, 240 msec). It was predicted that changes in vowel quality, with resulting changes in the extent of the formant transitions into and/or out of the closure period, would affect the single-cluster boundary more than the single-geminate boundary. On the other hand, changes in vowel duration, which might change the perceived speaking rate, were expected to affect the single-geminate boundary more than the single-cluster boundary. The data strongly support the first prediction and weakly support the second. They encourage the view that the two perceptual boundaries reflect the integration of temporally distributed information at two different levels of representation, one preceding and the other following phonetic categorization. The results are discussed in the light of three hypotheses about the role of silence in phonetic perception: silence as mere processing time for preceding cues (the backward masking hypothesis, from which the original predictions were derived), silence as the only cue (the differentiation hypothesis), and silence as information about articulatory gestures (the articulatory hypothesis).

## INTRODUCTION

Silence plays an important role in the perception of stop consonants. As its name implies, the articulation of a stop consonant involves an interruption of the airflow from the lungs by briefly closing the vocal tract. Thus, silence (or near-silence) is the most characteristic acoustic feature of stop consonants embedded in an utterance. This is reflected in perception. Under certain conditions, the presence of silence can be a sufficient cue for hearing a stop consonant. For example, by artificially introducing a silent interval between the fricative S-noise and the vocalic LIT-portion of the word

**Acknowledgment:** This research was supported by NICHD Grant HD01994 and BRS Grant RR05596. A summary of the results was presented at the 95th Meeting of the Acoustical Society of America in Providence, Rhode Island, May 1978. Helpful comments on an earlier draft were provided by Leigh Lisker, Richard Pastore, and David Pisoni.

SLIT, its perception can be changed to SPLIT (Bastian, Eimas, & Liberman, 1961). By inserting silence just before the SH-noise in SAY SHOP, this utterance can be converted into SAY CHOP, where the affricate CH is essentially a stop-initiated fricative (Dorman, Raphael, & Liberman, in press). Silence is not only sufficient but also necessary for the perception of stops in these contexts (i.e., between a fricative noise and a vocalic portion, in either order), since a naturally produced SPLIT or SAY CHOP can, in general, be turned into SLIT and SAY SHOP, respectively, by eliminating the silent closure period of the stop consonant from the speech signal (Repp, Liberman, Eccardt, & Pesetsky, 1978; Fitch, Erickson, Halwes, & Liberman, 1979).

Silence appears to be less crucial for the perception of stop manner in intervocalic position. In general, silence is neither sufficient nor necessary for the perception of stop manner in this context (i.e., between two vocalic portions), although it is, of course, an essential acoustic feature of natural productions. Thus, inserting silence between two steady-state vowels will, in general, not lead to a stop percept, and eliminating the closure period of a natural intervocalic stop consonant generally leaves the percept intact.[1] However, this does not imply that intervocalic silence plays no role in perception. For one, it may convey information about the voicing and even about the place of articulation of intervocalic stops (Lisker, 1957, 1978; Port, 1977). Its role in the perception of stop manner--i.e., in detecting the presence of an intervocalic stop consonant--is probably reduced by the general availability of other powerful manner cues: the transitions of the various formants (especially the first) into and especially out of the closure period. Although such transitions are characteristically also present in fricative-vowel and vowel-fricative contexts, they seem to carry greater perceptual weight in purely vocalic environments, thus decreasing the relative importance of silence.

Silence regains its primary perceptual importance when not one but a sequence of two different stop consonants is to be detected in vocalic context, as, for example, in the nonsense utterance /ebde/. When the closure period of a naturally produced utterance of this sort is spliced out, listeners typically hear only the second stop consonant but not the first, that is, /ede/. In order to perceive the first consonant, between 50 and 100 msec of silence is needed between the two vocalic portions, depending on the particular stimuli used (Repp, 1978; Dorman et al., in press). In natural speech, closure durations for sequences of two different stops are about twice as long as those for single stops--about 170 msec vs. 80 msec, on the average (Westbury, Note 1). Given that silence is indispensable for perception of the first stop in /ebde/ but not for perception of the second stop, and given that perception of the first stop breaks down when the natural closure duration is cut down to approximately half its size, it seems as if the first half of the closure "belonged to" the first stop and the second half to the second stop.[2]

According to one theoretical view (Repp et al., 1978; Dorman et al., in press), silence is a perceptual cue equivalent to other acoustic cues for stop manner, e.g., formant transitions. All these cues are supposed to provide information about the articulatory act that has produced the signal. An alternative view considers silence merely as time for undisturbed processing of the cues preceding it; thus, the perceptual disappearance of the first stop in /ebde/ after the closure has been eliminated would be an instance of recognition backward masking (Massaro, 1975, pp. 125-150; Repp, 1978). According to that interpretation, a certain amount of silence is required to fully process the cues contained in the first vocalic portion, which convey stop manner and, in particular, the place of articulation of the first consonant in the sequence. If the silence is too short, processing of that information will remain incomplete and will not lead to a separate phonemic percept. Repp (1978) has furnished some evidence that this incompletely processed information is not lost but integrated with the cues for the place of articulation of the second stop consonant. There is a third possible interpretation of the /ebde/ phenomenon--that the silent interval leads the listener to differentiate stimulus portions that otherwise would be processed as a single unit. These different hypotheses will be treated in more detail under General Discussion.

The listener's problem in the /ebde/ paradigm seems not to lie in the detection of manner information for the first stop consonant. If 50-100 msec of silence merely told the listener that two stop consonants have occurred (/ebde/), rather than one (/ede/), the same amount of silence presumably should signal the presence of two identical stop consonants (e.g., /ed-de/) when the cues on both sides of the closure period signal the same place of articulation. In fact, however, a much longer silent interval is needed to evoke the perception of double (i.e., geminate) stop consonants--approximately 200 msec in English-speaking listeners (Pickett & Decker, 1960; Repp, 1978). Thus, the silent interval required to perceive a sequence of two intervocalic stop consonants is much longer when the two phonemes are the same than when they differ in place of articulation.[3] This suggests that the silence in /ebde/ is required to interpret the place-of-articulation cues in the vocalic portion preceding the silence, and to determine that they result from a different place of occlusion than the cues in the vocalic portion following the silence.

Thus, we must distinguish two, different, critical silence durations in the perception of multiple intervocalic stop consonants; they will be referred to in the following as the single-cluster and single-geminate boundaries.[4] The _single-cluster boundary_ is the amount of silence needed to correctly identify a cluster of two different stop consonants, e.g., /ebde/, on 50 percent of the trials. It is obtained experimentally by varying the amount of silence between the syllables /eb/ and /de/ and by determining the point at which the probability of hearing /ebde/ equals that of hearing /ede/ (the perceptual boundary). The _single-geminate boundary_ is the amount of silence needed to perceive a sequence of two identical stop consonants, e.g., /ed-de/. It is obtained experimentally by varying the amount of silence between, say, the syllables /ed/ and /de/ and by determining the point at which the probability of hearing /ed-de/ equals that of hearing /ede/.

Repp (1978) hypothesized that the single-cluster and single-geminate boundaries reflect the average limits of two different processes of perceptual integration over time: The single-cluster boundary reflects the limit of the perceptual integration of place-of-articulation cues (primarily spectral in nature) occurring before and after the silence into a single phonemic percept, regardless of whether these cues signal different places of articulation or not. Such integration would occur as long as the pre-closure cues have not been fully processed at the phonetic level, and it would occur at a precategorical level of processing, since only a single phonemic decision is made on the basis of already integrated information. The single-geminate boundary, on the other hand, is assumed to reflect the limit of a higher-level integration process: The cues preceding the silent closure interval are phonetically interpreted (given that the silence exceeds the single-cluster boundary), and only if the phonetic interpretation of the cues following the closure yields an identical outcome are the two phonetic decisions integrated into a single phonemic percept (given that the silence does not exceed the single-geminate boundary). In other words, the single-geminate boundary is assumed to reflect a perceptual-phonological rule prescribing that two identical phonetic decisions merge into a single percept as long as their temporal separation does not exceed a certain limit.

The present experiments deal with two factors that were expected to have unequal effects--if any--on the single-cluster and single-geminate boundaries: spectral vs. temporal changes in the vocalic portions surrounding the silent interval. The basis for the prediction of unequal effects is the hypothesis--outlined above--that the single-cluster boundary reflects the processing and precategorical integration of spectral cues over time, whereas the single-geminate boundary reflects a perceptual-phonological rule that operates after the phonetic interpretation of acoustic cues. Thus, spectral changes in the vocalic portions and hence in the transitional spectral cues for stop manner and place, with resulting changes in their detectability or salience, should affect the single-cluster boundary more than the single-geminate boundary. On the other hand, it was thought that changes in the durations of the vocalic portions, and hence in prosodic features such as perceived stress and rate of articulation, might affect the single-geminate boundary more than the single-cluster boundary, especially since the former is known to be highly sensitive to variations in speaking rate (Pickett & Decker, 1960).

## EXPERIMENT I

In Experiment I the quality of the vowels ($V_1$, $V_2$) was manipulated preceding and following the silent closure period in utterances of the type /$V_1$b-g$V_2$/. Although a variety of different vocalic contexts has been used in past studies of the single-cluster boundary (Repp, 1978; Dorman et al., in press), no systematic comparison of different contexts has been made, and it is not known to which degree the boundary varies with different vowels. It is well-known, however, that different vocalic portions surrounding a stop closure interval exhibit quite different formant transitions for the same place of articulation. Depending on the characteristic formant frequencies of the vowel and the place of occlusion, the formant transitions may be more or less pronounced. It may be hypothesized that extensive formant movements provide stronger place and manner cues than shallow transitions, so that the first consonant in a cluster may be easier to detect (i.e., require a shorter processing time, and hence a shorter silent interval) when the transitions into the closure are large, than when they are small.

Conversely, there may also be an effect of the magnitude of the formant transitions following the silent interval. To the extent that the cues for the second stop in the cluster override or mask those for the first, and if the masking power of transitional cues depends on their salience and processing time, it may be predicted that larger transitions following the silence will shift the single-cluster boundary toward longer silence durations, whereas small transition cues for the second stop will make the first stop easier to detect and shift the boundary toward shorter silence durations. On the other hand, quite the opposite result would be predicted if the perceptual salience of the transitions on either side of the closure contributed to the _differentiation_ of the two sets of cues by serving as more effective boundaries of the silent interval. In this case, large formant transitions following the closure should lead to short single-cluster boundaries, and small transitions to long boundaries.

There are thus two contrasting predictions for the effects of spectral variations following the silence, whereas the potential effect of spectral variations preceding the silence appears to have only one reasonable predicted direction. In order to study the effects of both factors, they were varied orthogonally. The orthogonal design also made it possible to look for an interaction between the two factors. Such an interaction might be obtained if the relative continuity of formant trajectories across the silence, or any other relationship between the two sets of spectral cues, plays a perceptual role.

An analogous design was used to study the effect of spectral variations on the single-geminate boundary. However, it was expected that this boundary would depend primarily on the duration of the silent interval, and little or not at all on auditory factors, such as the extent of formant transitions. Such an outcome would be in agreement with the hypothesized higher-level, postcategorical nature of the single-geminate distinction.

## Method

_Subjects_. Twelve subjects participated. They included the author, a research assistant, and ten paid student volunteers with varying experience in listening to synthetic speech.

_Stimuli_. In the single-cluster condition, the vowel-consonant (VC) syllables /ib/, /ab/, /ub/ were followed by the consonant-vowel (CV) syllables /gi/, /ga/, /gu/ after varying intervals of silence. In the single-geminate condition, the same VC syllables were followed by /bi/, /ba/, /bu/. Thus, the consonants were kept unchanged within each task, and the initial and final vowels were varied orthogonally, resulting in nine combinations.

The stimuli were generated on the OVEIIIc serial resonance synthesizer at Haskins Laboratories. A set of natural utterances was converted into synthesizer parameters by means of a computer program (CONVERT).[5] The original utterances contained single stop consonants embedded between various vowels (e.g., /abi/, /uga/, etc.). The present stimuli consisted of the initial and final vocalic portions of several of these original utterances. In selecting the stimuli, note was taken of the fact that the formant transitions in a given vocalic segment did not depend to any significant extent on the vocalic segment beyond the stop closure, in this set of utterances at least. (This is contrary to Ohman's, 1966, data for Swedish.) Not knowing of any relevant evidence (such as a spectrographic study of formant movements in stop clusters), the present author assumed that the formant transitions would also be appropriate for stops occurring in clusters.

Schematic three-formant spectrograms and amplitude contours of the stimuli are shown in Figure 1. The contours in the figure correspond to the input parameters to the synthesizer, except that, for display purposes, they have been smoothed over the stepwise changes occurring every 10 msec, the time frame duration used in synthesis. In order to reduce uncontrolled acoustic variability, the original stimuli were "regularized" in various ways. All VC syllables were made 180 msec long by extending or cutting back the steady-state vowel at the beginning. All CV syllables were adjusted to 290 msec duration by similar manipulations of the steady state at the end. Irregularities in the steady-state formants due to frequency jitter in the CONVERT procedure were smoothed out, and uniform fundamental frequency contours (not shown in Figure 1) were imposed on all stimuli. Fundamental frequency was constant at 120 Hz in all VC syllables and during the the first 100 msec of CV syllables; it fell linearly from 120 to 100 Hz over the remaining 190 msec in CV syllables. Linearly rising and falling amplitude contours were imposed on the first half of VC stimuli and the second half of CV stimuli, respectively. Note that none of these changes affected the stimulus portions containing the formant transitions; they remained as traced from the original utterances. CV syllables beginning with /g/ showed a marked plateau in formants and amplitudes during the first 20 msec. Most likely, this portion marked the occurrence of a release burst in the original stimuli. However, the synthetic versions were created with periodic excitation throughout, without any apparent loss in intelligibility.

One additional regularizing procedure did affect the transitional portions. Due to the nature of the serial resonance synthesizer, the stimuli varied widely in amplitude: Stimuli with a high first formant (/a/) were considerably more intense than stimuli with a low first formant (/u/, /i/). Although such a difference mimics natural speech, to some extent at least, it seemed too large to be desirable in the present experiment. Therefore, the overall amplitudes of the stimuli were adjusted so as to give equal VU-meter readings across different vocalic contexts. These adjustments were performed on the digitized waveforms of the synthetic stimuli, using the pulse code modulation system at Haskins Laboratories with a 10kHz sampling rate. Figure 1 shows the adjustments in dB by the numbers below the amplitude contours; the contours themselves are shown as they were before the adjustment. (Note that the stimuli in each column of Figure 1 were adjusted only relative to each other.)[6]

Two stimulus tapes were recorded, one for each of the two conditions. Each tape contained four random sequences of 99 stimuli each. These 99 stimuli resulted from nine VC-CV combinations, each presented with eleven different closure (silence) durations. In the single-cluster condition (/ib/, /ab/, /ub/ followed by /gi/, /ga/, /gu/), the intervals ranged from 15 to 115 msec in 10-msec steps. In the single-geminate condition (/ib/, /ab/, /ub/ followed by /bi/, /ba/, /bu/), they ranged from 115 to 315 msec in 20-msec steps.[7] The interstimulus interval was 3 sec, and there were longer pauses between blocks of 99.

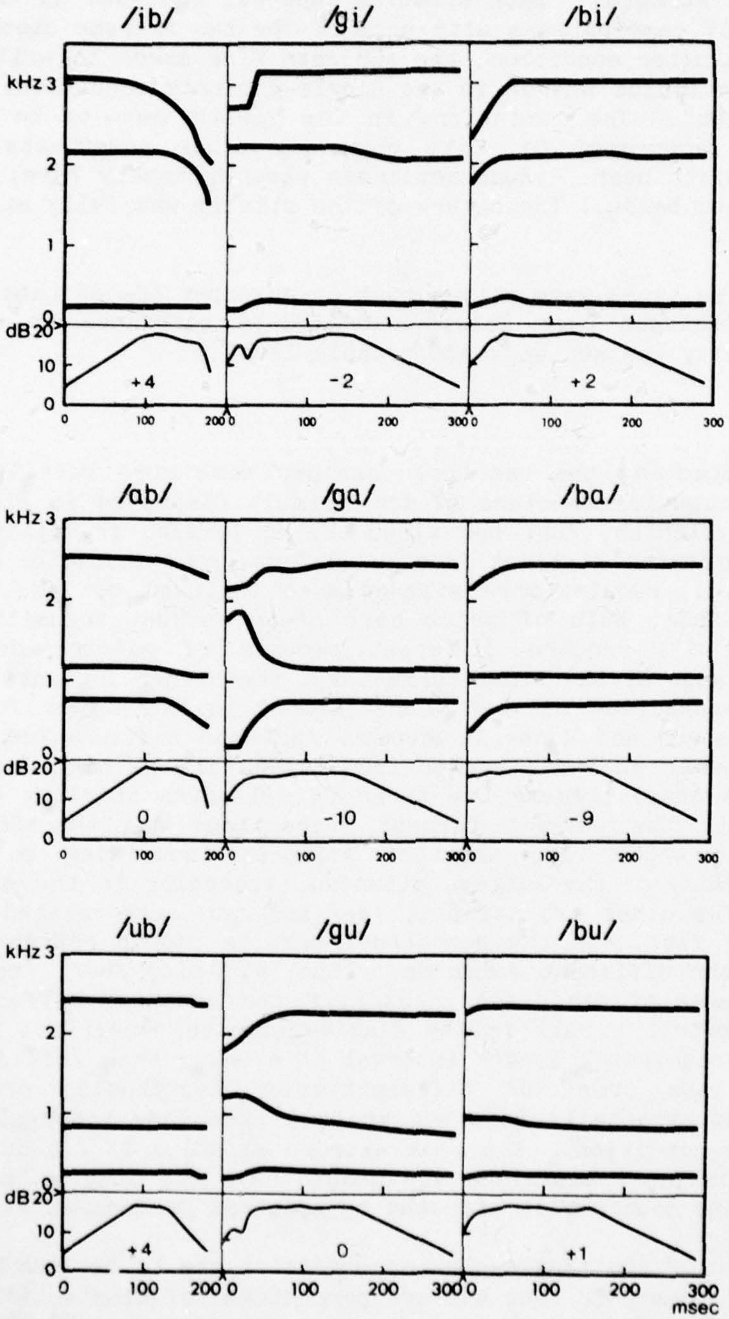


Figure 1. Schematic spectrograms and amplitude contours of the stimuli used in Experiment I. The numbers below the amplitude contours indicate later adjustments (dB) in overall amplitude, relative to the other stimuli in the same column (see text).

Procedure. Each subject participated in two one-hour sessions. In each session, both stimulus tapes were presented; their sequence was counterbalanced across subjects and reversed between the first and second sessions of each subject. Thus, each subject gave a total of 8 responses to each individual stimulus. Each stimulus tape was preceded by examples of the nine basic VC-CV combinations with each of the two extreme closure durations. In the single-cluster condition, the subjects were asked to write down "g" or "bg" for each stimulus heard; in the single-geminate condition, the responses were "b" and "bb". The variations in the vowels were to be ignored. The subjects were encouraged to write down any other consonants or consonant clusters they might hear. (Such responses were extremely rare; one exception will be mentioned below.) The nature of the stimuli was fully explained before the experiment.

The stimulus tapes were played back on an Ampex AG-500 tape recorder, and the subjects listened over Telephonics TDH-39 earphones in a quiet room. Playback intensity was set at a comfortable level.

## Results

Before inspecting the results, consider some more detailed predictions based on the acoustic structure of the stimuli displayed in Figure 1. Among the three VC syllables, /ub/ has virtually no formant transitions. (Indeed, /ub/ is only minimally distinct from /u/.) Thus, /ub/ should be more difficult to perceive--i.e., require more silence when followed by /gi/, /ga/, /gu/--than /ib/ and /ab/, both of which have clear formant transitions. Whether /ib/ and /ab/ will require different amounts of silence depends on the relative importance of the first formant vs. the higher formants: /ib/ has no first-formant transition but large transitions in the higher formants (which cue both stop place and manner), whereas /ab/ has a first-formant transition (primarily a manner cue) and weaker transitions in the higher formants (here primarily place cues). Among the three CV syllables starting with /g/, /gi/ is the one with the weakest formant transitions and therefore should be associated with either the shortest silences (according to the backward masking hypothesis) or the longest silences (according to the differentiation hypothesis). The other two stimuli, /ga/ and /gu/, are related to each other as are /ab/ and /ib/, and the perceptual results should reflect the relative importance of the different formants. Finally, /bi/, /ba/, /bu/ are more or less mirror images of /ib/, /ab/, /ub/; if the spectral differences between them have any effect at all in the single-geminate condition, they should be such that /bu/ requires a longer interval of silence than /bi/ or /ba/. This prediction follows from the differentiation hypothesis; presumably, the backward masking hypothesis does not apply at the long intervals used in the single-geminate condition. The most extreme stimulus is /ub-bu/; it contains virtually no formant transitions and should have the longest single-geminate boundary, if this boundary is affected by spectral factors at all.

The results of the single-cluster condition are to be found in the left-hand panels of Figure 2. The average percentages of cluster ("bg") responses are shown as a function of closure duration for the nine VC-CV combinations. Effects of the CV portions can be seen by comparing the three vertical panels, whereas effects of the VC portions are represented within each panel.

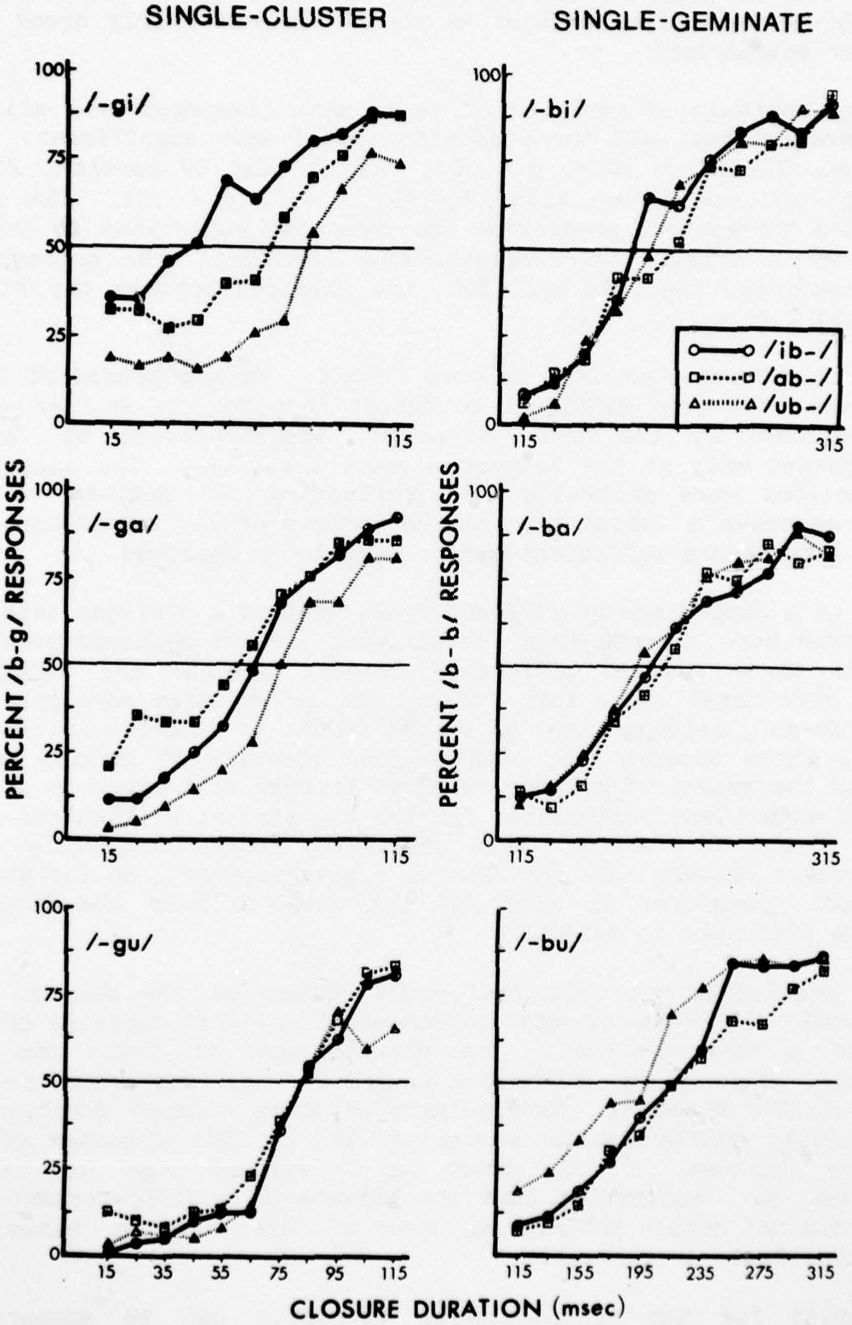


Figure 2. Average response percentages as a function of closure duration for the nine stimulus combinations in each of the two conditions of Experiment I.

The single-cluster response functions for all nine stimulus combinations rose as closure duration increased, reflecting the increasing probability of detecting the /b/. The functions were not very steep, which was in part due to considerable variation in criteria between subjects. As predicted, there were substantial differences between the various response functions; they crossed the 50-percent lines between 40 and 80 msec--a fairly broad range of single-cluster boundaries.

A two-way analysis of variance of these data (averaged over all closure durations) revealed that all three effects tested were significant: that of the VC portion, F(2,22) = 10.4, $\underline{p} < .01$; that of the CV portion, F(2,22) = 4.9, $\underline{p} < .05$; and the interaction, F(4,44) = 4.6, $\underline{p} < .01$. The two main effects confirm the general prediction that spectral variations in the VC and CV portions would affect the single-cluster boundary. The presence of an interaction indicates that, in addition, the relation between the VC and CV portions played a role.

Now let us examine the data in more detail. It was predicted that the /b/ in /ub/ would be more difficult to detect than the /b/ in /ib/ and /ab/. This was confirmed by the data, although, when followed by /gu/, the difference emerged only at the longest closure durations. The asymptotes of the /ub/-functions were generally low, reflecting the failure of several listeners to perceive a labial stop in the absence of any significant formant transitions, even when a sufficient amount of silence followed.

Turning to a comparison of /ib/ and /ab/, we find a striking interaction: /ab-gi/ required more silence than /ib-gi/, but /ib-ga/ required more silence than /ab-ga/; there was no difference between /ib-gu/ and /ab-gu/. In addition, we have noted above that /ub-gu/ did not require more silence than /ib-gu/ and /ab-gu/, although the /b/ was difficult to perceive in /ub-ga/ and /ub-gi/. All three observations suggest that identity of initial and final vowels reduced the amount of silence required to hear both stops in a cluster, and that this effect was responsible for the significant statistical interaction between VC and CV effects. This unexpected interaction fully accounts for the difference between /ib/ and /ab/ in a given context, so the absence of a first-formant transition in /ib/ did not seem to make the labial stop consonant more difficult to detect.

Another prediction was that /gi/ would either be the weakest masking stimulus or would afford the poorest differentiation, with opposite effects on the amount of silence required. The results seem to favor the masking hypothesis (see, however, the Discussion section): /gi/ was associated with a short single-cluster boundary. Particularly at short closure durations, /gi/ did not completely obliterate the preceding /b/, so that a number of cluster responses were obtained. Of the other two CV stimuli, /gu/ led to longer boundaries than /ga/, suggesting that the absence of a first-formant transition in /gu/ did not reduce the masking power of this stimulus (assuming that the masking hypothesis is correct).

The results for the single-cluster condition can be summarized as follows: (1) Among the "targets," /ub/ required more silence than /ib/ and /ab/, which differed little. (2) The effectiveness of the "maskers" increased from /gi/ to /ga/ to /gu/. (3) Less silence was required when the initial and final vowels were the same.[8]

We turn now to the results for the single-geminate condition, shown in the right-hand panels of Figure 2. The graphic display is analogous to that for the single-cluster condition, except for the appropriate changes in closure durations, CV portions, and responses. It is immediately evident that the single-geminate response functions showed smaller differences than the single-cluster functions. All average single-geminate boundaries (50-percent cross-overs) occurred in the range between 175 and 210 msec. Although this range is comparable to that obtained for single-cluster boundaries, it is small relative to the range of closure durations employed, which was twice as wide in the single-geminate condition.

A two-way analysis of variance of these data revealed no significant effects. The largest effect was the VC by CV interaction, $F(4,44) = 2.4$, $\underline{p} > .05$. Thus, the results bear out the prediction that spectral variations would not significantly affect the single-geminate boundary.[9] It is especially interesting to note that /ub/ suffered no perceptual disadvantage whatsoever in this task, despite its minimal formant transitions. The same is true for /bu/. The combination /ub-bu/ actually had a shorter single-geminate boundary than several other VC-CV stimuli, as can be seen in the lower right-hand panel of Figure 2.

## Discussion

The results of this experiment confirm the two general predictions made at the outset: The location of the single-cluster boundary is affected by spectral variations in the vocalic portions preceding and following the silence, whereas the single-geminate boundary is not significantly affected by such variations. These findings are consistent with the hypothesis that, at intervals shorter than the single-cluster boundary (i.e., shorter than 40-80 msec), the phonetic processing of the pre-closure place-of-articulation cues is disrupted by the early arrival of the post-closure cues, and that the time required to fully process the pre-closure cues (i.e., to categorize them phonetically and, thus, to perceive them as a separate phoneme) depends on their spectral saliency (i.e., extent of formant movements). At intervals between the single-cluster and the single-geminate boundary (80-175 msec), on the other hand, the processing of the pre-closure cues presumably can be completed within the closure interval. Thus, pre- and post-closure cues are categorized separately (though not independently; see Repp, 1978) and only subsequently integrated into a single phonemic percept, given that they signal the same place of articulation. This postcategorical integration then depends only on the interval elapsing between the two vocalic portions, not on the spectral properties of the sounds.

In the perception of stop clusters, the spectral properties of both the VC and the CV portion play a role. A stop consonant cued by minimal formant transitions into the closure period (/ub/) is difficult to detect, so that more silence is needed for its perception. A stop consonant cued by weak transitions out of the closure period (/gi/) is less effective as a masking stimulus, presumably because its processing takes longer to reach the stage of interference, so that less silence is needed for the perception of the first stop in the cluster. This latter result seems to refute an alternative interpretation: that more pronounced transitions increase the perceptual differentiation of the two vocalic portions, and thus of the two stop consonants. However, inspection of the waveforms of the stimuli revealed that

/gi/, the CV stimulus with the weakest transitions, had a burst-like onset, due to the closeness of its second and third formants (whose amplitudes interact in the serial synthesizer used). The CV syllable /ga/ had a weaker burst at onset, and /gu/, the most effective masking stimulus, had none at all and its first pitch pulse was so weak that its effective onset of energy was actually delayed by 8 msec (one pitch period). To the extent that the burst-like onset in /gi/ compensated for the absence of strong formant transitions, it may have led to better perceptual differentiation of the two vocalic portions in a VC-/gi/ stimulus, and thus to a reduction in the silence needed to hear the first stop. The burst-like onset of /gi/ may have acted as a better delimiter of the silent interval, thus increasing the effectiveness of silence as a stimulus separator or as a cue to stop manner in its own right (Dorman et al., in press). It seems quite likely that the effective duration of a silent interval would be longer when its boundaries are less distinct. This must have been especially true in the case of /gu/, whose apparent masking power most likely was due to its weak first pitch pulse. Thus, the data seem to be compatible with both the backward masking and the differentiation hypothesis.

Perhaps the most intriguing result is the finding that less silence was needed to perceive clusters when the surrounding vowels were the same than when they were different. This effect is interesting because it suggests a possible articulatory correlate: It seems plausible that the additional articulatory maneuvers required to change the shape of the vocal tract from one vowel to another may increase the closure duration of naturally produced intervocalic stop clusters. Of course, if a case is to be made for a close tie between perception and production in the single-cluster distinction (Dorman et al., in press), the total patterns of the present results should be paralleled by systematic variations in closure durations as a function of vocalic contexts in (rapidly spoken) natural utterances. A study investigating this issue is planned. However, the effects observed may also have a purely psychoacoustic explanation. This issue will be taken up again in the General Discussion.

## EXPERIMENT II

The second experiment examined the effects of a different factor on the single-cluster and single-geminate boundaries: the durations of the initial and final vocalic portions. In a design similar to that of Experiment I, these durations were varied orthogonally, so that their independent effects could be evaluated, as well as their interaction. Variations in the durations, $D_1$ and $D_2$, of the two vocalic portions have two different effects on the prosodic characteristics of the utterance: They change the perceived stress (a function of $D_1 - D_2$) and the perceived rate of articulation (presumably a function of $D_1 + D_2$). Thus, perceived stress depends on the relative durations of the two vocalic portions, whereas perceived speaking rate is likely to depend on the total duration of the utterance. Each effect implies independent perceptual roles of VC duration and CV duration.

It is not known to what extent perceived stress might influence either the single-cluster or the single-geminate boundary; therefore, the predictions in the present experiment were less clear than in Experiment I. However, earlier results provide some basis for predicting that perceived speaking rate might affect the single-geminate boundary more than the single-cluster boundary. Pickett and Decker (1960) have shown that the single-geminate boundary is highly sensitive to the rate of articulation of a sentence frame; the boundary may shift over a range as wide as 150 msec (150-300 msec). To the extent that durational changes in isolated disyllabic utterances can actually convey to the listener changes in rate of articulation (and this is by no means certain), similar shifts should be found in the present experiment; that is, the single-geminate boundary should become longer as the total duration of the stimuli increases. On the other hand, there is some evidence in the literature that the perception (and production) of silent closure intervals is less influenced by speaking rate in the case of single intervocalic stops or affricates. Gay (1978) and Isenberg (1978) have shown that closure intervals change less than the surrounding vocalic portions with rate of articulation, though they do change. Recent perceptual results of Marcus (1978) and Repp et al. (1978) suggest that the perception of silent intervals is relatively insensitive to variations in speaking rate. Port (1977, 1978), on the other hand, demonstrated small but consistent speaking rate effects on the perception of intervocalic silence as a voicing and place cue. All these results were obtained with single stops in vowel or fricative-vowel environment, and it is not known whether they can be generalized to stop clusters. The weakest prediction one could make is that effects of stimulus duration on the single-cluster boundary should be small because of the relatively short silence durations involved.

## Method

Subjects. Eleven subjects participated, eight of whom (including the author) also were subjects in Experiment I. The three new subjects were paid volunteers from the same subject pool.

Stimuli. The stimuli /ib/, /ga/, and /ba/ were selected from the set used in Experiment I. The duration of /ib/ was modified by either adding or deleting steady-state vowel portions at the beginning. The durations of /ga/ and /ba/ were changed by deleting portions of the vowel at the end; this reduced the extent of the fundamental frequency contour, leading to a perceptual enhancement of the durational differences, since vowels with a pitch contour are perceived as longer than vowels without such a contour (Lehiste, 1976). The durations used were 120, 180, and 240 msec for both VC and CV stimuli.

The experimental tapes were exactly analogous to those in Experiment I, except that the orthogonal variations in duration took the place of the variations in vocalic context.

Procedure. The procedure was the same as in Experiment I.

## Results

The most consistent finding was that, surprisingly, CV duration had no systematic effect in either condition, nor were there any significant interactions between VC duration and CV duration. Thus, contrary to the hypotheses stated in the introduction to Experiment II, it was neither the difference between (or perhaps the ratio of) the VC and CV durations (stress) nor their sum (rate of articulation) that affected the two perceptual boundaries.

Rather, only the duration of the VC portion played a role. Therefore, the results were averaged over the CV duration factor. The effect of VC duration in each condition is shown in Figure 3; the top panels show the group results (N = 11), whereas the bottom panels show the data for a single experienced listener (BHR), as discussed below.

The effect of VC duration was significant both in the single-cluster condition, $F(2,20) = 8.7$, $\underline{p} < .01$, and in the single-geminate condition, $F(2,20) = 3.5$, $\underline{p} = .05$. As expected, the amount of silence at the perceptual boundary increased with VC duration. The second effect was at least as large as the ıirst; its marginal significance level was due to high inter-subject variability in the single-geminate condition. In particular, two of the eleven subjects showed an effect of VC duration on the single-geminate boundary that was opposite in direction to the effect shown by the other subjects. When the data of these two subjects were removed, the effect of VC duration on single-geminate judgments increased both in average magnitude and reliability, $F(2,16) = 11.5$, $\underline{p} < .01$.[10] Figure 3 shows the average data of all subjects, however.

In general, single-geminate judgments were difficult to make in the presence of random variations in stimulus duration, as indicated by the very flat response functions. Single-cluster judgments, on the other hand, were barely disturbed by durational variability. (Compare the slopes of the functions in Figure 2 with those in Figure 3.) This provides some indirect (and unexpected) evidence in support of the hypothesis that the single-geminate distinction would be more affected by variations in stimulus duration than the single-cluster distinction.

Since the author was by far the most experienced subject and, in fact, the only listener to produce clean data in the single-geminate condition, his results deserve special attention. They are shown in the bottom panels of Figure 3. As in the group data, only VC duration had a systematic effect. However, this effect was clearly more pronounced in the single-geminate condition than in the single-cluster condition, as originally predicted. Thus, while the group data are equivocal, the results for this single experienced listener support the predictions made at the outset.[11]

## Discussion

The results of Experiment II show that both perceptual distinctions investigated were affected by changes in stimulus duration, and that the stimulus portion preceding the critical silent closure interval was solely responsible for this effect. The average effect of VC duration on the single-geminate boundary was about twice as large as that on the single-cluster boundary. This may have been due to the simple fact that the critical closure durations in the former condition were about twice as long as in the latter. However, even larger quantitative differences were observed for several subjects, including the author. In addition, qualitative differences between the VC duration effects in the two conditions were suggested by the general disturbing effect of durational variations on single-geminate judgments but not on single-cluster judgments. There were also somewhat different patterns of results in the two conditions; cf. the different spacing of the response functions in Figure 3, top panels. However, the evidence from the group data remains merely suggestive of different effects of VC duration on the two

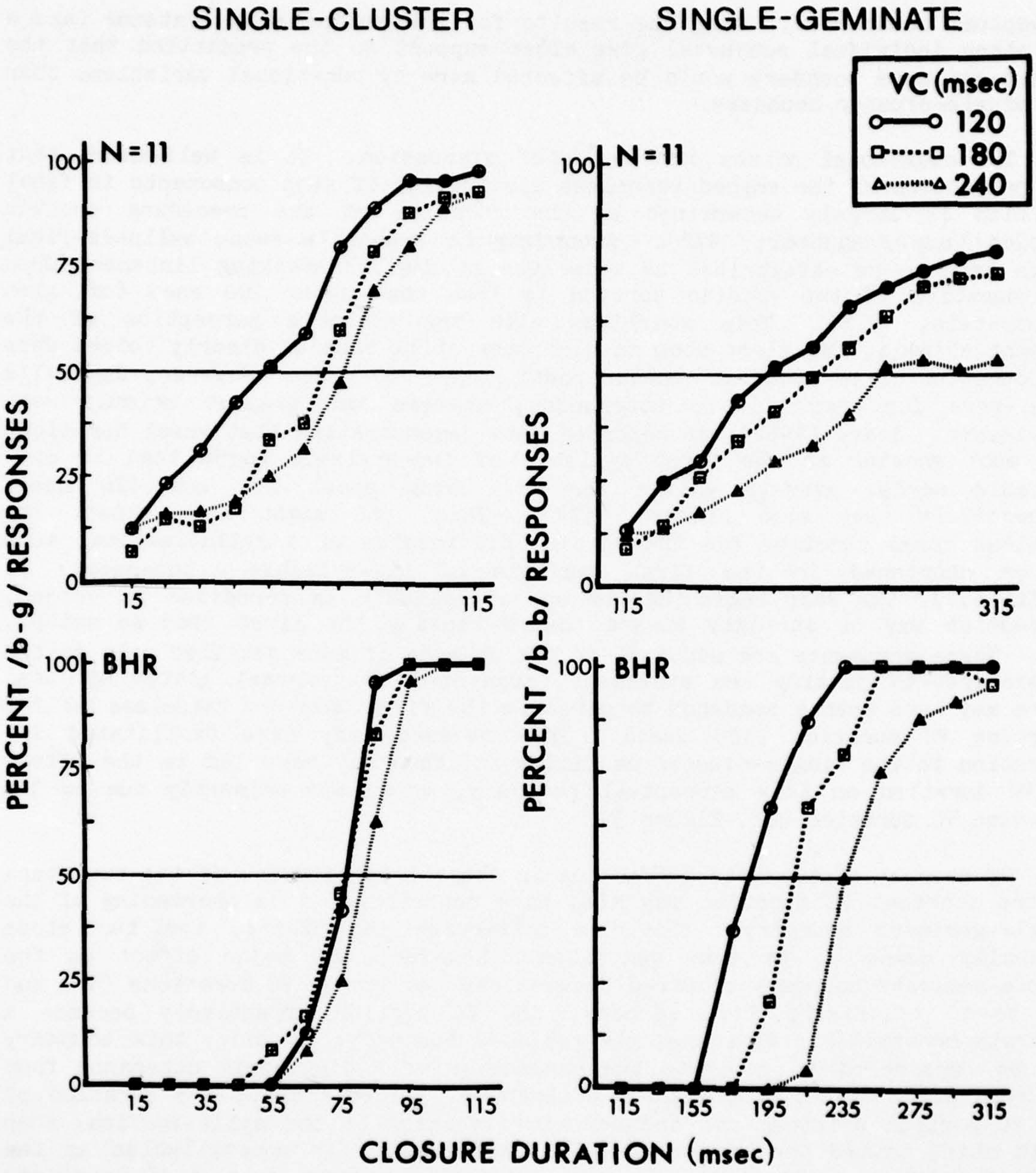


Figure 3. Effect of the duration of the preceding vocalic portion on the single-cluster and single-geminate distinctions. Average response percentages for a group of 11 subjects (top) and for a single experienced listener (bottom).

perceptual boundaries. Only the results for the author as a listener (and a few other individual subjects) give clear support to the prediction that the single-geminate boundary would be affected more by durational variations than the single-cluster boundary.

Two additional points deserve brief discussion. It is well-known that the perception of the voiced-voiceless distinction of stop consonants in final position is largely determined by the duration of the preceding vocalic portion (e.g., Raphael, 1972). According to Raphael's data, syllable-final stops tend to be categorized as voiceless by English-speaking listeners when the duration of the vocalic portion is less than about 200 msec (cf. also Mermelstein, 1978). This conflicts with the author's perception of the present stimuli; the first stop in a cluster still sounded clearly voiced when the duration of the initial vocalic portion was 180 msec. However, Raphael's data were for monosyllabic utterances, whereas the present stimuli were disyllabic. Sharf (1962) has reported data demonstrating that vowel durations are much shorter in the first syllable of two-syllable words than in one-syllable words, average values for /i/ being about 200 and 120 msec, respectively (see also Lisker, 1974). Thus, one might also expect the critical vowel duration for the voicing distinction of a syllable-final stop to be shortened in the first syllable of two-syllable utterances. In addition, if the stop beginning the second syllable is perceived as voiced, perception may be strongly biased toward hearing the first stop as voiced, too. These arguments are adduced--in the absence of more detailed data in the literature--to justify the statement, suggested by informal listening, that there may have been a tendency to perceive the first stop as voiceless at the shortest VC duration (120 msec). This tendency may have facilitated its detection in the single-cluster condition and thus may have led to the effect of VC duration on this perceptual boundary, which was primarily due to the 120-msec VC duration (cf. Figure 3).

Of course, a perceived difference in the voicing feature of the two stops at the shortest VC duration may also have contributed to a shortening of the single-geminate boundary. (Given a difference in voicing, the two stops naturally cease to be true geminates.) However, the major effect on the single-geminate boundary occurred between the two longer VC durations (180 and 240 msec, cf. Figure 3). Perhaps, the VC portion effectively becomes a separate monosyllable at the single-geminate boundary; in fact, this boundary may be considered to separate the perception of a disyllabic utterance from the perception of two monosyllabic utterances. In this case, the duration of the VC portion critical for the perceived voicing of the syllable-final stop would shift toward the longer duration appropriate for monosyllables as the silence duration approaches the single-geminate boundary, so that VC durations in the vicinity of 200 msec would have the largest perceptual effects. Thus, both single-cluster and single-geminate results may reflect the effects of VC duration as a voicing cue. The main problem with this explanation is that, in natural articulation, closure durations tend to be longer when the first stop in a cluster is voiceless than when it is voiced (Westbury, Note 1), so that it is difficult to see why a tendency to perceive the first stop as voiceless should have reduced the silence required for its perception in a cluster.

A second point to be made concerns the nature of the information conveying rate of articulation. The initial assumption was that it is the total stimulus duration that carries this information. However, recent experiments by Summerfield (Note 2) indicate that the perceptual effects of speaking rate are not mediated by parameters extracted over longer stretches of the signal, but rather are primarily due to the temporal properties of the acoustic segment immediately preceding the critical cues. The present data clearly lend themselves to this interpretation. The effects of VC duration thus may legitimately be considered effects of perceived speaking rate; more appropriately, perhaps, they should be considered instances of a perceptual effect of "local" temporal properties of the signal, of which those effects often thought to be due to speaking rate are an instance.

A recent investigation by Port (1978) further supports this interpretation. He showed that a presumed speaking rate effect on closure duration as a voicing cue for the intervocalic stop in RABID-RAPID (Port, 1977) was primarily due to changes in duration of the first syllable in the test word, and that the voicing boundary (i.e., the critical closure duration) changed in proportion to the duration of this syllable. Such proportionality was not observed in the present data, possibly due the fact that, in contrast to Port's stimuli, the duration of the first syllable was not a direct cue for the phonetic distinctions investigated.

In summary, VC duration may have had two different kinds of perceptual effects in the present experiment--one on the perceived voicing of the syllable-final stop, and one on the effective duration of the silent closure period. However, it is not known to which extent either of these effects contributed to the observed overall effect of VC duration on the single-cluster and single-geminate boundaries.

## GENERAL DISCUSSION: THREE HYPOTHESES CONCERNING THE ROLE OF SILENCE IN SPEECH PERCEPTION

The present experiments have yielded several interesting empirical findings. These findings were discussed and, in part, had been predicted within the framework of several alternative hypotheses concerning the role of silence in speech perception. The experiments were not designed to provide a critical comparative test among these hypotheses, and, indeed, they do not strongly favor one over another. One reason it is difficult to make strong theoretical statements at this point is that our knowledge about the psychoacoustics of speech perception on the one hand, and about speech production on the other, has seriously fallen behind the multiplicity of perceptual results. This final section summarizes the various hypotheses and points out what sort of information should be added to the present data in order to decide between the conceptual alternatives.

### The Backward Masking Hypothesis

This hypothesis, due to Massaro (1975, pp. 125-150), assumes that silence protects the processing of preceding information from interference by later information. Thus, the function of silent intervals in speech is to provide time for undisturbed processing--a hypothesis that deliberately ignores the articulatory function of such intervals. There are two versions of the hypothesis regarding the nature of the interference: It may be an all-or-none process, so that later cues interrupt the processing of earlier cues and the results of that processing are lost; or it may be gradual, so that whatever information has been extracted from the partially processed cues is integrated with the results of processing the more dominant, later cues. The former version seems to be implied by Massaro (1975, pp. 125-150), whereas the latter version has been favored by the author (see Repp, 1978). Of course, the distinction between interruption and integration hypotheses is familiar from the literature on visual backward masking (e.g., Scheerer, 1973).

The backward masking hypothesis applies only to the single-cluster distinction. At closure durations appropriate for testing the single-geminate distinction, the processing of the pre-closure cues presumably can be completed well before the end of the closure period. There is no reason why the processing of pre-closure cues should take longer when those cues represent the same place of articulation as do the post-closure cues. Therefore, a perceptual rule needs to be postulated that accounts for the combination of two fully processed cues for the same phoneme into a unitary percept, if these cues occur within a certain time span. Such a rule may be called perceptual-phonological because it operates on the output of the hypothetical phonetic processing stage. Given these assumptions, the single-cluster and single-geminate boundaries are seen to constitute the limits of perceptual integration at two different levels, one precategorical (i.e., preceding phonetic categorization) and the other postcategorical (Repp, 1978).

The results of Experiment I may be taken to support the backward masking hypothesis. Strong transitional cues for place of articulation presumably speed up processing, and therefore less silence is needed following such cues in the single-cluster condition. In the single-geminate condition, processing is completed within the closure interval no matter how weak the cues, and therefore no effect of spectral variations is observed. Similarly, if the cues following the silence are prominent, they will be processed faster and therefore reach the interference stage earlier, so that a longer silence is required to protect the processing of the pre-closure cues (interruption hypothesis). Alternatively, the relative strength of the post-closure cues may be manifested in the weight they receive when they are integrated with incompletely processed pre-closure information (integration hypothesis). In the single-geminate condition, no effect of spectral variations in post-closure cues is observed, simply because no interference takes place. Presumably, the single-geminate judgment is based on the duration of the silent interval alone. The lack of a significant effect of spectral variations in this condition suggests that the critical temporal cue is _not_ the interval between the hypothetical completion times for processing the two sets of cues; otherwise, the results should have paralleled those in the single-cluster condition. Rather, the temporal cue for the single-geminate distinction seems to be picked up at the level of early auditory analysis.

One result of Experiment I that is not explained by the backward masking hypothesis is the facilitative effect of identical vowels preceding and following the closure. Although effects of similarity between target and masker in auditory recognition masking have been observed (Kallman & Massaro, 1978), they have been in the direction that greater similarity led to more masking, not less. Massaro (1974), in particular, observed low performance when the vowel /a/ was used both as a target and as a mask in a random sequence including other vowel combinations. Thus, the present finding of facilitation in a superficially similar situation contradicts the backward masking hypothesis.

The results of Experiment II also create problems for the backward masking hypothesis. It will be recalled that the duration of the VC portion affected the single-cluster boundary (in the majority of subjects, at least), whereas the duration of the CV portion had no effect. The absence of a CV duration effect corresponds to an absence of a masker duration effect in auditory recognition backward masking (Massaro, 1971). To account for the VC duration effect, however, it must be assumed that an increase in the duration of the initial steady-state vowel slows down the processing of the following transitional cues. The opposite would be expected if, for example, an earlier onset of the VC portion reduced temporal uncertainty about the occurrence of the transition cues. Thus, the direction of the VC duration effect seems counterintuitive in the framework of the backward masking hypothesis.

In summary, although some of the present results fit the backward masking hypothesis, as formulated by Massaro (1975, pp. 125-150), others are not compatible with it. An elaborated form of the hypothesis that accounts for all the results may have to incorporate assumptions that are more germane to the two other hypotheses to be discussed.

## The Differentiation Hypothesis

The differentiation hypothesis (explicitly formulated here for the first time) attempts to account for the present findings by assuming that the listeners' judgments in the present tasks are based entirely on the perceived duration (or detectability) of the silent closure interval. If the interval is perceived as too short (or is not detected at all), the pre- and post-closure cues are assumed to be perceptually integrated, with the post-closure cues receiving higher weights in the process. (This assumption is shared with the integration version of the backward masking hypothesis.) If the interval is sufficiently long, differentiation (perceptual separation) of pre- and post-closure cues will occur, enabling the listener to perceive them as separate phonemes.

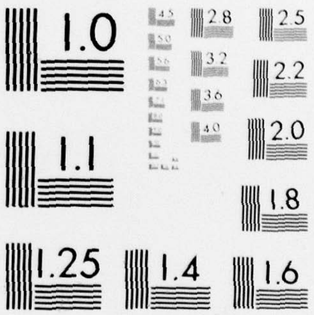
It may be (but need not be) the case that the single-cluster boundary represents a detection threshold for the silent interval. To explain why the single-geminate boundary is much longer than the single-cluster boundary, it is necessary to postulate different perceptual criteria in the two cases. Certainly, there is no reason for a silent interval to appear longer when the transitional cues on either side signal the same place of articulation. A separate criterion (or perceptual rule) for the single-geminate distinction also had to be postulated in the framework of the backward masking hypothesis. That hypothesis is not distinct from the differentiation hypothesis as far as the single-geminate distinction is concerned.

The differentiation hypothesis is difficult to evaluate because we know relatively little about the psychoacoustics of interval perception. For example, it is not even clear whether it is reasonable to assume a detection threshold for silence as long as 40-80 msec, since no studies of gap detection have employed interval markers as complex as the ones used here. It is known, however, that silent intervals are more difficult to detect in heterogeneous contexts than in a homogeneous medium (Perrott & Williams, 1971; Williams & Perrott, 1972; Collyer, 1974) and that duration discrimination suffers similarly in heterogeneous contexts (Divenyi & Danner, 1977). Thus, the differentiation hypothesis can explain why the single-cluster boundaries occurred at

AD-A083 179 HASKINS LABS INC NEW HAVEN CONN F/G 5/10
SPEECH RESEARCH, A REPORT ON THE STATUS AND PROGRESS OF STUDIES--ETC(U)
MAR 79 A S ABRAMSON, T BAER, C BEST, G CARDEN PHS-HD-01994
UNCLASSIFIED SR-57(1979) NL

4 OF 4
AD
A083179

END
DATE
FILMED
5-80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

shorter silences when the initial and final vowels were identical.

According to the differentiation hypothesis, all observed boundary shifts contingent upon variations in the VC and CV stimulus portions are due either to changes in the effectiveness with which these portions delimit the silent interval or to other, more indirect, effects they have on the perceived duration of that interval. In the discussion of Experiment I, it has been pointed out that the effects of varying vocalic context on the single-cluster boundary are compatible with such an explanation, although independent evidence from analogous nonspeech studies is needed. It is not clear, however, why the single-geminate boundary was not similarly affected by spectral stimulus variations. Perhaps, the longer interval durations and listener variability precluded such effects from appearing in the data, although this seems a weak excuse.

The results of Experiment II seem to be compatible with the differentiation hypothesis. In one of the few nonspeech studies in the literature directly relevant to the present data, Penner (1976) found that randomly varying the durations of noise markers impaired discrimination of silent interval durations, and that this impairment was primarily due to variations in the first marker. It remains to be seen, however, whether the subjective duration of silent intervals increases with the duration of the preceding marker, as the results of Experiment II suggest. Williams and Perrott (1972) found that the gap detection threshold increased when the durations of both markers were increased simultaneously. Further nonspeech experiments are needed to assess the usefulness of the differentiation hypothesis. At present, this hypothesis actually fits the data better than the backward masking hypothesis.

## The Articulatory Hypothesis

The third hypothesis deserving serious consideration is based on the idea that speech perception is guided by implicit knowledge of articulatory dynamics. Unlike the other two views--one of which considers silence as a mere vehicle of processing, whereas the other takes it to be the only cue--the articulatory hypothesis considers silence as a cue on a par with other cues resulting from the same underlying articulatory act (Repp et al., 1978; Dorman et al., in press). As with the two other hypotheses, the articulatory hypothesis needs to incorporate an additional assumption to explain why the pre-closure cues (and not the post-closure cues) suffer perceptually when closure duration is too short. It is not clear whether this assumption can be framed in articulatory concepts; alternative concepts, such as "weighted integration," may have to be borrowed.

One strong postulate of the articulatory hypothesis is that the single-cluster boundary should correspond to the minimal possible closure interval in articulation, and similarly, that the single-geminate boundary should reflect longer closure durations for this distinction in speech production. While there is some support for the second contention (Pickett & Decker, 1960; but see Footnote 3), the first one has yet to receive direct support. One basic problem is that perceptual results are likely to be sensitive to a variety of factors entirely unrelated to speech production, such as range, contrast, and other criterion effects. Therefore, any absolute comparison of perceptual data with acoustic speech measurements is highly problematic.

A relative comparison of patterns of results between perception and production holds more promise. Essentially, the articulatory hypothesis predicts that, for each perceptual effect obtained, there is a corresponding variation in the durations of naturally produced closure intervals. Again, this prediction is probably exaggerated; there may be certain perceptual effects that are psychoacoustic in origin and therefore have no articulatory correlate (cf. Repp, in press, for an example). Although the extent to which perception parallels production will become clear only after the relevant acoustic analyses of natural utterances have been conducted, there are some perceptual effects that are particularly suggestive of perception-production correlations. The best example is the shortening of the single-cluster boundary observed in homogeneous vocalic contexts (Experiment I). There is preliminary evidence from acoustic measurements that shorter closure durations may indeed be associated with such contexts.

In summary, it may well be true that none of the individual hypotheses discussed above is alone sufficient to account for the perceptual results. Although the articulatory hypothesis is an attractive notion, particularly with regard to speech perception in natural situations, its applicability is likely to be limited by the very disturbances we create by manipulating speech experimentally. It may prove necessary to borrow concepts from the other hypotheses described (and perhaps others not even mentioned) to account for a number of specific findings due to psychoacoustic or methodological factors. The untangling of these (themselves not uninteresting) effects from those due to articulatory knowledge is one of the major problems faced by current speech perception research.

## REFERENCE NOTES

1. Westbury, J. R. Temporal control of medial stop consonant clusters in English. Paper presented at the 93rd Meeting of the Acoustical Society of America, State College, Pennsylvania, June 1977.
2. Summerfield, A. Q. On articulatory rate and perceptual constancy in phonetic perception. Unpublished manuscript, 1977.

## REFERENCES

Abbs, M. H. A study of cues for the identification of voiced stop consonants in intervocalic contexts. Unpublished Ph.D. dissertation, University of Wisconsin, 1971.

Bastian, J., Eimas, P. D., & Liberman, A. M. Identification and discrimination of a phonemic contrast induced by silent interval. Journal of the Acoustical Society of America, 1961, 33, 842. (Abstract)

Collyer, C. E. The detection of a temporal gap between two disparate stimuli. Perception & Psychophysics, 1974, 16, 96-100.

Divenyi, P. L., & Danner, W. F. Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. Perception & Psychophysics, 1977, 21, 125-142.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, in press.

Fitch, H. L., Erickson, D., Halwes, T. G., & Liberman, A. M. A trading relation in perception between silence and spectrum. Haskins Laboratories Status Report on Speech Research, 1979, SR-57.

Gay, T. Effect of speaking rate on vowel formant movements. Journal of the Acoustical Society of America, 1978, 63, 223-230.

Isenberg, D. Relative duration of stop closure and fricative noise across speaking rate. Journal of the Acoustical Society of America, 1978, 63 (Supplement No. 1), S54-55. (Abstract)

Kallman, H. J., & Massaro, D. W. Similarity effects in backward recognition masking. WHIPP Report No. 4 (Department of Psychology, University of Wisconsin), 1978.

Lehiste, I. Influence of fundamental frequency pattern on the perception of duration. Journal of Phonetics, 1976, 4, 113-117.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. Language, 1957, 33, 42-49.

Lisker, L. On "explaining" vowel duration variation. Glossa, 1974, 8, 233-246.

Lisker, L. Closure hiatus: Cue to voicing, manner, and place of consonant occlusion. Haskins Laboratories Status Report on Speech Research, 1978, SR-53, vol. 2, 79-86.

Marcus, S. M. Distinguishing "slit" and "split"--an invariant timing cue in speech perception. Perception & Psychophysics, 1978, 23, 58-60.

Massaro, D. W. Effect of masking tone duration on preperceptual auditory images. Journal of Experimental Psychology, 1971, 87, 146-148.

Massaro, D. W. Perceptual units in speech perception. Journal of Experimental Psychology, 1974, 102, 199-208.

Massaro, D. W. Perceptual images, processing time, and perceptual units in speech perception. In D. W. Massaro (Ed.), Understanding language. An information-processing analysis of speech perception, reading, and psycholinguistics. New York: Academic, 1975.

Mermelstein, P. On the relationship between vowel and consonant identification when cued by the same acoustic information. Perception & Psychophysics, 1978, 23, 331-336.

Ohman, S. E. G. Coarticulation in VCV utterances. Journal of the Acoustical Society of America, 1966, 39, 151-168.

Penner, M. J. The effect of marker variability on the discrimination of temporal intervals. Perception & Psychophysics, 1976, 19, 466-469.

Perrott, D. R., & Williams, K. N. Auditory temporal resolution: Gap detection as a function of interpulse frequency disparity. Psychonomic Science, 1971, 25, 73-74.

Pickett, J. M., & Decker, L. R. Time factors in the perception of a double consonant. Language and Speech, 1960, 3, 11-17.

Port, R. F. The influence of tempo on stop closure duration as a cue for voicing and place. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 59-74.

Port, R. F. Effects of word-internal versus word-external tempo on the voicing boundary for medial stop closure. Journal of the Acoustical Society of America, 1978, 63 (Supplement No. 1), S20. (Abstract.)

Raphael, L. J. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. Journal of the Acoustical Society of America, 1972, 51, 1296-1303.

Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Perception & Psychophysics, 1978, 24, 471-485.

Repp, B. H. Relative amplitude of aspiration as a voicing cue for syllable-initial stop consonants. *Language and Speech*, in press.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of temporal cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 621-637.

Scheerer, E. Integration, interruption and processing rate in visual backward masking. *Psychologische Forschung*, 1973, *36*, 71-93.

Sharf, D. J. Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech*, 1962, *5*, 26-30.

Williams, K. N., & Perrott, R. R. Temporal resolution of tonal pulses. *Journal of the Acoustical Society of America*, 1972, *51*, 644-647.

## FOOTNOTES

[1] These statements are based on informal observations, since no systematic investigations are available. (However, see Abbs, 1971.) Vowels with a very low first formant perhaps constitute an exception; for example, /u/-silence-/u/ may be heard as /ubu/ or /upu/, and the stop in a natural /ubu/ may disappear if the closure period is spliced out.

[2] This seems a reasonable interpretation, since natural productions of stop sequences usually exhibit a release burst (at the place of articulation of the first stop) that divides the closure period into two parts, the first of which then is naturally part of the acoustic information for the first consonant. Studies on the role of silence in stop sequence perception generally have not included release bursts; rather, bursts have been spliced out from natural speech and omitted in synthetic utterances. It is likely that the perception of the first stop in a sequence would remain unimpaired if the pre-release closure were left intact and only the silence following the release (perhaps even including the release burst itself) were removed.

[3] An exception to this result, when naturally produced geminates are used, has recently been found by Dorman, Raphael, and Isenberg (personal communication).

[4] Strictly speaking, neither the term "cluster" nor the term "geminate" is fully appropriate for sequences of stops crossing a syllable boundary. However, in the absence of any better mnemonics, their use in the present context seems justifiable.

[5] The CONVERT program enables the user to trace (with a light-pen) formants, fundamental frequency, and amplitude contours in a spectrographic display. The tracings become input to the synthesizer, thus creating a rather faithful, though simplified, copy of the original. Thanks are due to Gary Kuhn for his permission to make use of these stimuli. The conversion was done by him for a different purpose.

[6] In hindsight, these amplitude manipulations, even though they were meant to create more equal conditions, constitute a complicating factor in the experiment. An experiment is now in progress to determine whether amplitude variations affect either of the two perceptual boundaries. The present results revealed no perceptual disadvantages (in the form of more silence required) for /ab/, /ga/, and /ba/, whose amplitude had been sharply reduced.

[7] The intervals had been intended to range from 0-100 msec and from 100-300 msec, respectively, but account had to be taken of the fact that the synthesis parameters for the VC stimuli contained a "zero frame" at the end (a common procedure to avoid clicks), so that 10 msec of silence followed the digitized waveforms. The digitizing procedure itself introduced a 4-msec silent interval at the beginning of the CV stimuli. Thus, the actual intervals were 14 msec longer than intended (1 msec shorter than described in the text).

[8] It was gratifying to note that the author--the only experienced listener in the group--showed a pattern of results very similar to (but considerably less variable than) that of the group as a whole.

[9] Although there was no overall effect, a few individual subjects showed idiosyncratic differences. The most striking of these occurred for the author as a listener, since he tended to misperceive /bi/ as /di/. (This curious bias was not shared by any other subject.) Naturally, then, he showed shorter boundaries in combinations including /bi/, since he perceived them as containing /b-d/ clusters, not geminates. Interestingly, however, he did perceive single labial (not alveolar) stops at the shortest closure durations used (100-140 msec). Thus, he heard the /bi/ correctly given that it was not too far apart from the preceding labial VC formant transitions--an instance of perceptual integration of information across the closure interval (cf. Repp, 1978). Exclusion of this subject's data from the group data would not have altered any of the conclusions drawn from them.

[10] The cause for the reversal of the effect of VC duration for these two subjects is not known. One of the two anomalous subjects also provided curious data in the single-cluster condition (and in earlier experiments as well), but the other one seemed to be a fairly reliable listener. There was a third listener who produced very messy results only in the single-cluster condition. After removal of the two poorest subjects in that condition, the effect of VC duration on the single-cluster boundary increased in reliability, too, $F(2,16) = 15.1$, $\underline{p} < .01$, without changing the basic pattern of the data.

[11] The totally random stimulus sequences precluded any possible influence of the author's hypotheses on his perception of the stimuli. Moreover, he took pains not to listen in an analytic fashion.

## II. PUBLICATIONS

## III. APPENDIX

# PUBLICATIONS

Bell-Berti, F., Baer, T., Harris, K. S., & Niimi, S. Coarticulatory effects of vowel quality on velar function. *Phonetica*, in press.

Bell-Berti, F., & Harris, K. S. Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America*, 1979, *65*, 771-773.

Best, C. T., & Harris, L. J. Childhood development. In *Encyclopedia of Psychology*. Princeton, N.J.: Arétè Press, in press.

Fujimura, O., Baer, T., & Niimi, S. A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation. *Journal of the Acoustical Society of America*, 1979, *65*, 478-480.

Harris, K. S. Vowel duration change and its underlying mechanisms. *Language and Speech*, in press.

Healy, A. F. Poor communication in psycholinguistics: Review of four new textbooks. *Journal of Psycholinguistic Research*, 1978, *7*, 477-492.

Liberman, A. M., & Pisoni, D. B. Evidence for a special speech-perceiving subsystem in the human. In T. H. Bullock (Ed.), *The recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen, 1977.

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz, & H. L. Teuber (Eds.), *Handbook of sensory physiology*, Vol. VII, "Perception." Heidelberg: Springer-Verlag, 1977.

Mann, V. A., Hein, A., & Diamond, R. Localization of targets by strabismic subjects: Contrasting patterns in constant and alternating suppressors. *Perception & Psychophysics*, 1979, *25*, 29-34.

Mann, V. A., Hein, A., & Diamond, R. Patterns of interocular transfer of visuomotor coordination reveal differences in the representation of visual space. *Perception & Psychophysics*, 1979, *25*, 35-41.

McGarr, N. S., & Osberger, M. J. Pitch deviancy and intelligibility of deaf speech. *Journal of Communication Disorders*, 1978, *11*, 237-247.

Miller, J. L., & Liberman, A. M. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, in press.

Raphael, L. J., Bell-Berti, F., Collier, R., & Baer, T. Tongue position in rounded and unrounded front vowel pairs. *Language and Speech*, 1979, *22*, 37-48.

Repp, B. H. Accessing phonetic information during perceptual integration of temporally distributed cues. *Journal of Phonetics*, in press.

Repp, B. H. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, in press.

Repp, B. H., Healy, A. F., & Crowder, R. G. Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 129-145.

# APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-57:

| Status Report | | DDC | ERIC |
|---|---|---|---|
| SR-21/22 | January - June 1970 | AD 719382 | ED-044-679 |
| SR-23 | July - September 1970 | AD 723586 | ED-052-654 |
| SR-24 | October - December 1970 | AD 727616 | ED-052-653 |
| SR-25/26 | January - June 1971 | AD 730013 | ED-056-560 |
| SR-27 | July - September 1971 | AD 749339 | ED-071-533 |
| SR-28 | October - December 1971 | AD 742140 | ED-061-837 |
| SR-29/30 | January - June 1972 | AD 750001 | ED-071-484 |
| SR-31/32 | July - December 1972 | AD 757954 | ED-077-285 |
| SR-33 | January - March 1973 | AD 762373 | ED-081-263 |
| SR-34 | April - June 1973 | AD 766178 | ED-081-295 |
| SR-35/36 | July - December 1973 | AD 774799 | ED-094-444 |
| SR-37/38 | January - June 1974 | AD 783548 | ED-094-445 |
| SR-39/40 | July - December 1974 | AD A007342 | ED-102-633 |
| SR-41 | January - March 1975 | AD A013325 | ED-109-722 |
| SR-42/43 | April - September 1975 | AD A018369 | ED-117-770 |
| SR-44 | October - December 1975 | AD A023059 | ED-119-273 |
| SR-45/46 | January - June 1976 | AD A026196 | ED-123-678 |
| SR-47 | July - September 1976 | AD A031789 | ED-128-870 |
| SR-48 | October - December 1976 | AD A036735 | ED-135-028 |
| SR-49 | January - March 1977 | AD A041460 | ED-141-864 |
| SR-50 | April - June 1977 | AD A044820 | ED-144-138 |
| SR-51/52 | July - December 1977 | AD A049215 | ED-147-892 |
| SR-53 | January - March 1978 | AD A055853 | ED-155-760 |
| SR-54 | April - June 1978 | AD A067070 | ED-161-096 |
| SR-55/56 | July - December 1978 | AD A065575 | ** |
| SR-57 | January - March 1979 | ** | ** |

**DDC and/or ERIC order numbers not yet assigned.

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

UNCLASSIFIED

Security Classification

# DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>Haskins Laboratories<br>270 Crown Street<br>New Haven, Connecticut 06510 | 2a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED<br>2b. GROUP<br>N/A |

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 57, January-March, 1979

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

| | | |
|---|---|---|
| 6. REPORT DATE<br>March 1979 | 7a. TOTAL NO. OF PAGES<br>298 | 7b. NO. OF REFS<br>334 |
| 8a. CONTRACT OR GRANT NO.<br>HD-01994 NS13617<br>N01-HD-1-2420 AM25814<br>RR-5596<br>BNS76-82023<br>MCS76-81034<br>NS13870 | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>SR-57 (1979)<br>9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)<br>None | |

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited*

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

| | |
|---|---|
| 11. SUPPLEMENTARY NOTES<br>N/A contents: | 12. SPONSORING MILITARY ACTIVITY<br>See No. 8 |

13. ABSTRACT

This report (1 January - 31 March) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-An Articulatory Synthesizer for Perceptual Research;
-Perception of an Oral-Nasal Continuum Generated by Articulatory Synthesis
-Coarticulation and Theories of Extrinsic Timing
-Orthography and the Beginning Reader
-Aspiration Amplitude as a Voicing Cue for Syllable-Initial Stop Consonants Presented Monaurally and in Dichotic Competition
-Stop Consonant Place Perception with Single-Formant Stimuli: Evidence for the Role of the Front-Cavity Resonance
-Vowel Duration Change and its Underlying Physiological Mechanisms
-The Psycholinguistic Basis of Linguistic Awareness
-Some Effects of Later-Occurring Information on the Perception of Stop Consonant and Semivowel
-Perceptual Equivalence of Two Acoustic Cues for Stop-Consonant Manner
-Syllabic Coding and Reading Ability in Word Recognition
-Acoustic Cues for a Fricative-Affricate Contrast in Word-Final Position
-Apprehending Spelling Patterns for Vowels; A Developmental Study
"Perceptual Centers" in Speech Production and Perception
-Influence of Vocalic Environment on Perception of Silence in Speech.

DD FORM 1 NOV 65 1473 (PAGE 1)

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED

Security Classification

S/N 0101-807-6811

A-31408

| 14 KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Articulation | | | | | | |
|    Synthesis | | | | | | |
|    Perception | | | | | | |
|    Resonance, front cavity | | | | | | |
|    Vowel Duration | | | | | | |
| Coarticulation | | | | | | |
|    Timing | | | | | | |
| Perception | | | | | | |
|    Oral-nasal continua | | | | | | |
|    Syllable centers | | | | | | |
|    Silence, vocalic environment influence | | | | | | |
|    Place of articulation, stops | | | | | | |
|    Consonants | | | | | | |
|       Stops | | | | | | |
|       Voicing, syllable initial | | | | | | |
|       Contrasts, stop/semivowel initial | | | | | | |
|       Manner | | | | | | |
|    Contrasts, fricative/affricate | | | | | | |
| Reading | | | | | | |
|    Orthography | | | | | | |
|       Vowel patterns, developmental | | | | | | |
|       Beginning reader | | | | | | |
|    Psycholinguistic basis, linguistic awareness | | | | | | |
|    Syllabic coding, word recognition | | | | | | |

DATE

ILMED

-8